environmental microbiology reports

Environmental Microbiology Reports (2019) **11**(5), 672–689 doi:10.1111/1758-2229.12780

Brief Report

Prevalence of viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA

Carlos A. Ruiz-Perez,[1] Despina Tsementzi,[2] Janet K. Hatt,[2] Matthew B. Sullivan[4] and Konstantinos T. Konstantinidis[1,2,3*]

[1]*School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA.*

[2]*School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA.*

[3]*Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA, USA.*

[4]*Departments of Microbiology and Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH, USA.*

Summary

Bacteriophages encode host-acquired functional genes known as auxiliary metabolic genes (AMGs). Photosynthesis AMGs are commonly found in marine cyanobacteria-infecting *Myoviridae* and *Podoviridae* cyanophages, but their ecology remains understudied in freshwater environments. To advance knowledge of this issue, we analysed viral metagenomes collected in the summertime for four years from five lakes and two estuarine locations interconnected by the Chattahoochee River, Southeast USA. Sequences representing ten different AMGs were recovered and found to be prevalent in all sites. Most freshwater AMGs were 10-fold less abundant than estuarine and marine AMGs and were encoded by novel *Myoviridae* and *Podoviridae* cyanophage genera. Notably, several of the corresponding viral genomes showed endemism to a specific province along the river. This translated into *psbA* gene phylogenetic clustering patterns that matched a marine vs. freshwater origin indicating that *psbA* may serve as a robust classification and source-tracking biomarker. Genomes classified in a novel viral lineage represented by isolate S-EIVI contained *psbA*, which is unprecedented for this lineage. Collectively, our findings indicated that the acquisition of photosynthesis AMGs is a widespread strategy used by cyanophages in aquatic ecosystems, and further indicated the existence of viral provinces in which certain viral species and/or genotypes are locally abundant.

Received 26 January, 2019; accepted 29 June, 2019 *For corresponding. E-mail kostas@ce.gatech.edu. Tel. 404-385-3628; Fax 404-894-8266.

Introduction

Due to their vast numbers, phages can predate and control bacterial populations, and thus have a significant effect on global biogeochemical cycles (Weinbauer, 2004; Sime-Ngando and Colombet, 2009). However, their impact on bacteria is not restricted to population control by means of cell lysis (Suttle, 2002; Muhling *et al*., 2005), but also includes the mobilization and maintenance of host functional genes known as auxiliary metabolic genes (AMGs) (Thompson *et al*., 2011; Xia *et al*., 2013; Crummett *et al*., 2016). AMGs found in phage genomes encode proteins involved in various cellular functions including nucleotide synthesis/metabolism; carbon, nitrogen and sulphur metabolism; phosphate stress; cell protection and photosynthesis (Sullivan *et al*., 2010; Thompson *et al*., 2011; Crummett *et al*., 2016; Gao *et al*., 2016; Roux *et al*., 2016). Phage AMGs are expressed at different stages during infection (Lindell *et al*., 2005; Lindell *et al*., 2007). Their activity enhances key steps in bacterial metabolic pathways that are limiting during infection, presumably directing the host metabolism towards improved viral-particle production (Hurwitz and U'Ren, 2016; Breitbart *et al*., 2018; Fernandez *et al*., 2018).

AMGs involved in photosynthesis have attracted special attention due to the complexity of the photosynthetic machinery and the fitness advantages they could provide to cyanophages (Puxty *et al*., 2015). For instance, the photosystem II D1 and D2 protein-encoding genes (*psbA* and *psbD*) transcribed and translated during infection (Lindell *et al*., 2005; Sharon *et al*., 2007) have been reported to be widespread in marine viruses (Sullivan *et al*., 2006; Sharon *et al*., 2007; Chenard and Suttle, 2008). The photosynthesis-related gene repertoire found in viral genomes also includes genes involved in electron transfer (*psaA*, *petE*, *petF*, *ptoX*, *speD* and *hli*) and light harvesting processes (*ho1*, *pcyA*, *pebS* and *cpeT*) (Puxty *et al*., 2015; Crummett *et al*., 2016;

Gao *et al*., 2016). Given that cyanophages depend on host photosynthesis for their replication (Mackenzie and Haselkorn, 1972; Sherman, 1976; Gao *et al*., 2016), production of (viral) photosynthesis proteins during infection could compensate for the reduction in the active host transcript and protein pool leading to an increase in viral burst sizes. Therefore, the presence of such genes most likely represents a fitness advantage to cyanophages carrying them (Bragg and Chisholm, 2008; Hellweger, 2009; Gao *et al*., 2016).

Previous studies also suggested that phage-encoded PsbA in combination with other phage-encoded photosynthesis-related proteins such as high-light inducible proteins and ferredoxins might have played an important role in functionally differentiating *Prochlorococcus* subpopulations in marine systems (Lindell *et al*., 2004). For instance, phages have been shown to mediate the horizontal gene transfer (HGT) of *psb* genes between different bacterial lineages, directly affecting the evolution and diversity of these genes in their hosts (Sullivan *et al*., 2006). Nonetheless, the distribution of photosynthetic AMGs is not universal in all cyanophages, with most of them being present almost exclusively in *Myoviridae* genomes. *Podoviridae* representatives, on the other hand, usually encode only PsbA and high-light inducible proteins (Lindell *et al*., 2004; Zheng *et al*., 2013; Puxty *et al*., 2015), while there is only one report of a *Siphoviridae* representative encoding PsbA (Puxty *et al*., 2015). It has been hypothesized that the absence of photosynthesis-related genes in some cyanophages might be related to short latent periods of phage infection, under which circumstances a phage-encoded PsbA would not offer any advantage (Sullivan *et al*., 2006). However, more data are clearly needed to fully test this hypothesis.

Despite the knowledge gained about the diversity and function of viruses in the ocean, viral ecology in freshwater systems remains largely understudied. Although cyanophages from all three viral families *Myoviridae*, *Siphoviridae*, and *Podoviridae* were first recovered in freshwater ecosystems (Safferman and Morris, 1963; Xia *et al*., 2013), they have received less attention compared with marine cyanophages (Middelboe *et al*., 2008; Wilhelm and Matteson, 2008). Despite the recent increase in the number of studies targeting freshwater viral communities, there are still only about a dozen phage representatives with complete genome sequences (Liu *et al*., 2007; Liu *et al*., 2008; Dreher *et al*., 2011; Xia *et al*., 2013). Furthermore, while several phage representatives that infect freshwater cyanobacteria such as *Anacystis*, *Microcystis*, *Chroococcus*, *Aphanothece* and *Synechococcus* (unicellular), *Phormidium*, *Lyngbya*, *Plectonema*, *Anabaena*, *Planktothrix*, and *Nostoc* (filamentous) have been described (n = ~90) (Deng and Hayes, 2008; Sarma, 2012), only two of them carry photosynthesis-related AMGs. Specifically, the tailless phage PaV-LD isolated from the filamentous cyanobacterium *Planktothrix agardhii* encodes a non-bleaching protein A (NblA) involved in the degradation of phycobilisomes (Gao *et al*., 2012). Furthermore, the *Synechococcus*-isolated cyanophage S-CRM01 (*Myoviridae*), the first sequenced freshwater representative to carry the *psbA* gene, appears to be more similar to marine *Myoviridae* cyanophages (Dreher *et al*., 2011). Collectively, these previous findings indicated that phage-encoded photosynthetic AMGs might not be as widespread in freshwater ecosystems compared with their marine counterparts.

Few studies have discussed the presence of photosynthesis-related viral genes in freshwater lake systems using targeted PCR or DGGE and clone libraries, and virtually none has employed deep metagenomic sequencing to reconstruct the genomic context, ecology and prevalence of these genes. In these studies, however, the *psbA* gene has received increased attention given that it is widespread among marine *Myoviridae* and *Podoviridae* viral families, and thus could potentially serve as a better universal marker for cyanophages than the more restricted genes for capsid assembly protein-encoding gene *g20* and the DNA polymerase encoding gene (Gao *et al*., 2016). Culture-independent studies in Lake Erie (MI, USA), Lake Constance (Germany), Lake Annecy and Bourget (France) (Wilhelm *et al*., 2006; Chenard and Suttle, 2008; Wilhelm and Matteson, 2008; Zhong and Jacquet, 2013), and phage isolates from Lake Erie (MC15 and MC19) (Wilhelm *et al*., 2006) and the Klamath River in Northern California (S-CRM01) (Dreher *et al*., 2011) revealed that freshwater *psbA* gene sequences are, for the most part, evolutionarily divergent from their marine counterparts and that there is some level of separation between freshwater and marine *psbA* gene sequences. However, a larger representation of freshwater *Podoviridae* and *Myoviridae* genomes is necessary to fully validate the clustering patterns observed previously, especially across a (continuous) spatial freshwater-to-saltwater gradient (Zhong and Jacquet, 2013). Moreover, samples from rice paddy fields in Japan and China show intermixing between both environments and a separation between paddy water and other aquatic habitats (Wang *et al*., 2009; Wang *et al*., 2016) further complicating the evolutionary relationships among *psbA* sequences from different origins and stressing the need for genomic context and classification to resolve such discrepancies. Unfortunately, information about a *psbA*-encoding genome is only available for S-CRM01, limiting a more detailed overview of the much-needed genomic context in which this gene occurs in freshwater ecosystems.

To overcome the aforementioned limitations and provide new insight into the diversity and distribution of *psbAD* (and other photosynthesis-related) genes in freshwater ecosystems, we performed a comprehensive survey of cyanophage genomes encoding photosynthesis-

related genes using deep metagenomic sequencing on summertime samples originating from a four-year survey of five interconnected lakes and two estuarine locations along the Chattahoochee River, a major riverine ecosystem in Southeast United States. Our analyses provide a quantitative view of the presence, diversity, evolutionary history and genomic context of photosynthesis-related phage AMGs, and especially of the widespread photosystem II D1 *psbA* gene.

Results

Prevalence of viral genomes encoding photosynthesis AMGs along the Chattahoochee River

Viral particle purification and virome sequencing of samples from five lake sites interconnected by the Chattahoochee River in Southeastern USA, i.e., Lakes Lanier (LL), West-Point (LWP), Harding (LH), Eufaula (LE) and Seminole (LS) and two estuarine locations, Apalachicola Beach (APA) and East Point (E2M) (Fig. 1), yielded 20 viromes that were evaluated for bacterial contamination. Five viromes had 16S rDNA gene encoding reads in a proportion greater than 0.02%, an indication of the presence of bacterial-derived sequences and were therefore excluded from read-based AMG analyses (Supporting Information Fig. S1). The viromes were assembled and the viral origin of AMGs recovered in the assembled contigs was verified by manual inspection of the gene annotation, alignment and searching against the NCBI GenBank database for a viral gene best match. In total, 963 viral contigs containing at least one photosynthesis AMG were recovered (representing ~1% of the total sequence space assembled and dereplicated). Contigs were dereplicated at >95% identity along >80% of the length of the sequence resulting in a total of 783 unique genome fragments representing approximately species-level taxonomy (Brum *et al*., 2015; Gregory *et al*., 2016).

Among the dereplicated genome fragments, 153 represented long (>10 kbp) sequences, of which five were presumed to be complete due to assembling into circular genomes. Given that polyamines might be involved in non-photosynthesis functions (Wortham *et al*., 2007; Igarashi and Kashiwagi, 2010), viral genomes for which only *speD* (necessary for polyamine production) was detected were removed from further consideration to avoid including non-cyanobacterial infecting phages. After this step, 666 genomes were retained (79 longer than 10 kbp). Considering that viruses lack a common gene marker to assess classification and taxonomy, protein-sharing networks have emerged as an alternative method to assess viral taxonomy (Roux *et al*., 2016; Bolduc *et al*., 2017). Genome classification using proteins shared between the recovered and reference viral genomes by vConTACT2 resulted in a total of 362 viral clusters (VCs, approximating genus level) of which four contained exclusively genome fragments recovered in this study (Supporting Information **Fig. S2** and Tables S1 and **S2**). Out of the 79 long genomes, 15 clustered into a group that included reference *Podoviridae* T7-like phages from the RefSeq database, and thus were classifiable in this family (Supporting Information **Fig. S2**). The remaining 64 genomes with no representative genomes in the cluster were classified at least at the family level as *Podoviridae* and *Myoviridae* based on the edges they shared with the closest database representatives (Supporting Information **Fig. S2** and Tables S1 and **S2**). Interestingly, the cyanomyovirus group was formed by genome fragments recovered mostly from lake locations while the cyanopodovirus group included, almost exclusively, genome fragments from the estuarine locations. Notably, two genomes (VC_1010_5 and VC_1011_18) clustered with a novel lineage of *Synechococcus*-infecting freshwater cyanophages, S-EIVI (Chenard *et al*., 2015); an in-depth genome analysis revealed these three genomes share most genomic features except for the presence of *psbA* in the Chattahoochee River genomes. Overall, MetaVir2 annotations were consistent with those obtained using vConTACT2, when available (34/79). Our attempt to link some of our Chattahoochee River viral genomes to a host genome from the companion bacterial metagenomes were generally met with poor success (Supporting Information **Table S3**). Only one metagenome-assembled genome (MAG), classified in the *Synechococcaceae* family, was linked to a viral genome based on a CRISPR signature; the remaining 20 (of 79 in total) viral genomes were linked to non-cyanobacterial genomes, indicating that these were likely false-positive calls due to inaccuracies of the linkage approach or miss-assembly/miss-binning of the CRISPR- or tRNA-encoding contigs within the corresponding MAG (Supporting Information **Table S3**). Nonetheless, the abundance patterns of the Synechococcaceae host decrease when the phage is abundant and vice versa, suggesting a possible predator-pray dynamic (Supporting Information **Fig. S3**).

Estimated genome abundance profiles of the long genomes (n = 79) across the different sampling locations during late summer 2014–2015 were used to identify endemicity to specific regions (Fig. 2). The detected endemism was mostly driven by the differentiation between habitats, with 27 genomes showing a strong preference for estuarine locations and 34 for the freshwater lake locations; the remaining genomes showed no endemism or were detected too infrequently to infer endemicity patterns. We could identify three main groups within the freshwater-endemic group. The largest group (n = 16) showed endemism with most freshwater lakes, i.e., Lake Westpoint, Harding and Eufaula, while the two remaining groups showed strong endemicity with Lake Lanier (n = 8) and Lake

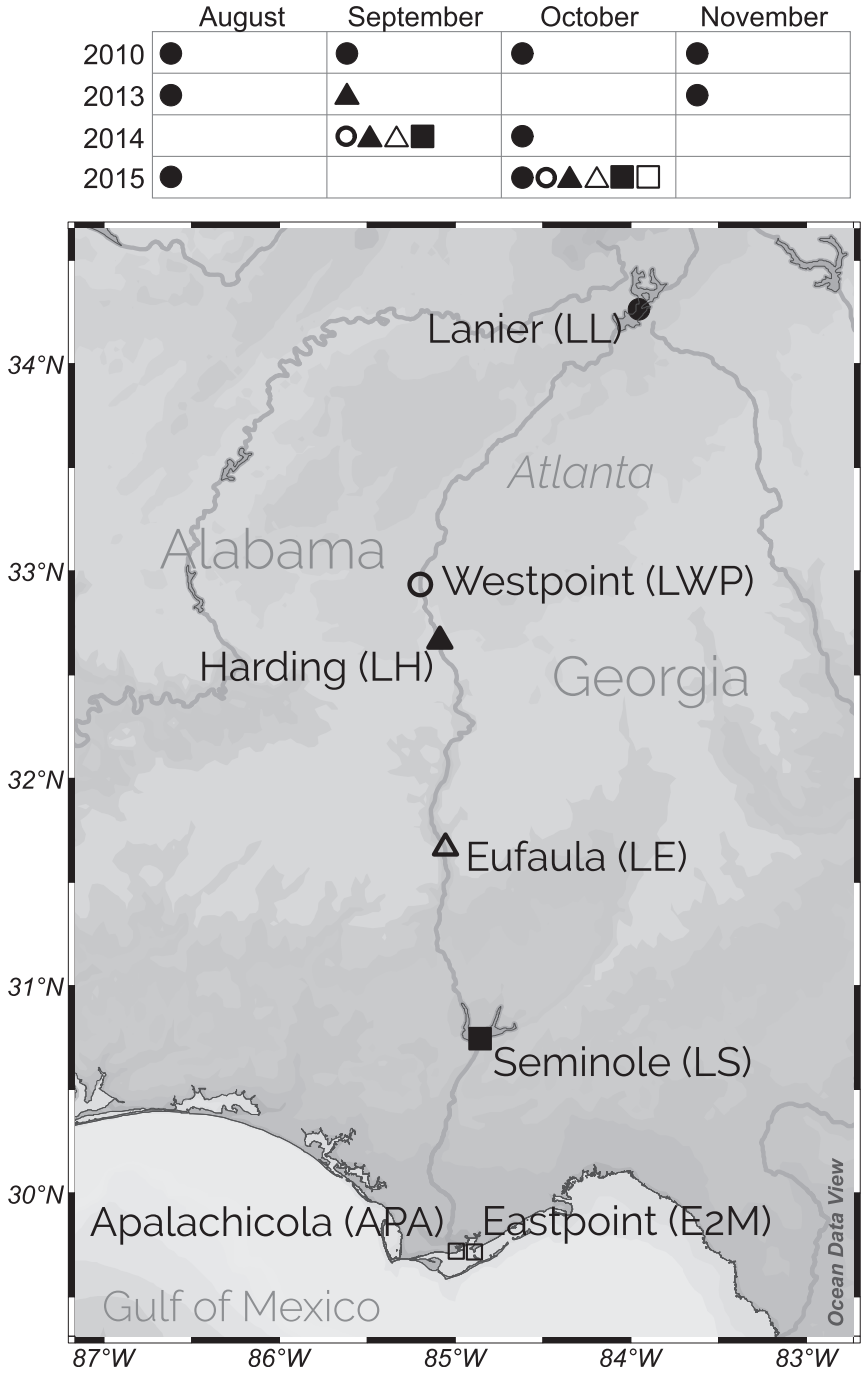

Fig. 1. Sample locations and dates of freshwater viral metagenomes.
Filled and unfilled shapes indicate the different sampling locations along the Chattahoochee River as five lake locations, Lake Lanier (LL); Westpoint Lake (LWP); Lake Harding (LH); Lake Eufaula (LE); Lake Seminole (LS), and two estuaries indicated by the same unfilled square, Apalachicola (APA) and Eastpoint (E2M). The sampling dates for each location are shown in the table above. Both estuarine samples were collected on the same date (October 2015).

Seminole (n = 10). Interestingly, all but one of the *Podoviridae* genomes (n = 15) were endemic exclusively to the estuaries, with just 15LWP_79 being endemic in multiple lakes, i.e., West Point, Harding and Eufaula. On the contrary, *Myoviridae* genomes were distributed to all endemism groups, i.e., were freshwater- or estuary-endemic.

Distribution of photosynthesis-related AMGs in different viral families and environments

The presence of photosynthesis-related AMGs, i.e., *psbA*, *psbD*, *psaA*, *speD*, *hli*, *petF*, *petE*, *ptoX*, *ho1*, *pebS*, *pcyA*, *cpeT* and *nblA* in the identified viral contigs was examined. The most prevalent photosynthesis-related AMG was *speD*

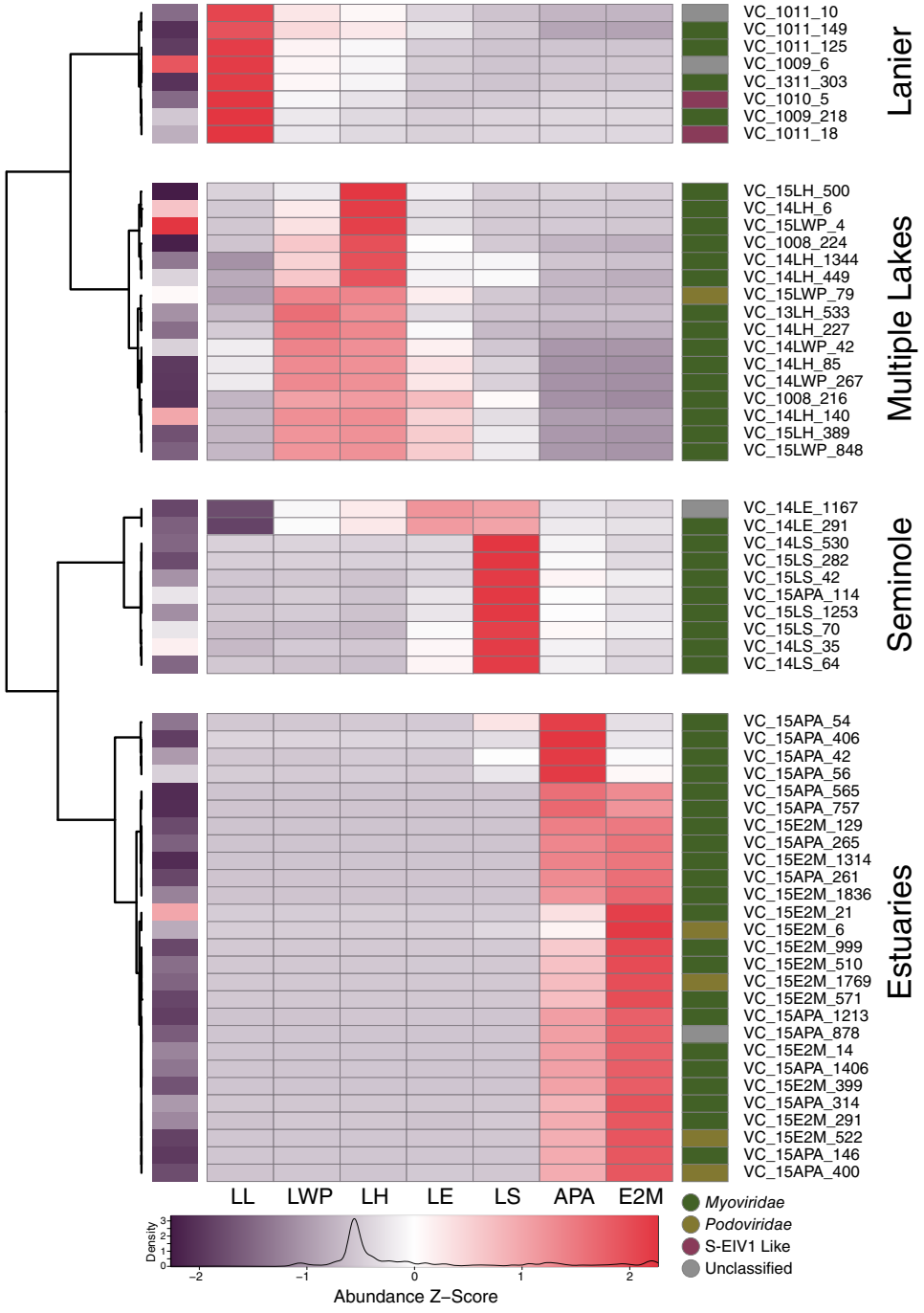

Fig. 2. Abundance profiles of endemic viral genome fragments along the Chattahoochee River basin.
Cubic smoothed splines were estimated with respect to fluvial distance from the most upstream Lake Lanier to the estuarine samples in late summer of 2014 and 2015. Genome fragments showing endemism were identified by a Pearson's R threshold of >0.5 between observed data and cubic splines. The predicted abundance per sampling location (columns) of each endemic genome fragment (rows) is displayed as log-abundances in Z-scores per row (see legend). Hierarchical clustering using correlation distances and the Ward method was used to separate between groups associated with Lake Lanier (LL), with multiple lakes (West Point, LWP; Harding, LH; Eufaula, LE), with Lake Seminole (LS) and finally with the estuaries (APA and E2M).

(∼34%), which became only 4% when excluding genomes encoding only *speD* (discussed above), and was followed by genes encoding high-light inducible proteins present in ∼28.5% of the genome fragments. Although *psbA* and *psbD* were not the most abundant AMGs recovered, they were found in a substantial portion of the contigs (27.5% and 9.45%, respectively). The remaining AMGs, besides *psbA*, *psbD*, *hli* and *speD* genes, were present in ∼10% of the genomes, except for *ho1*, *pebS* and *cpeT*, which were present in less than 2% of the genomes. In general,

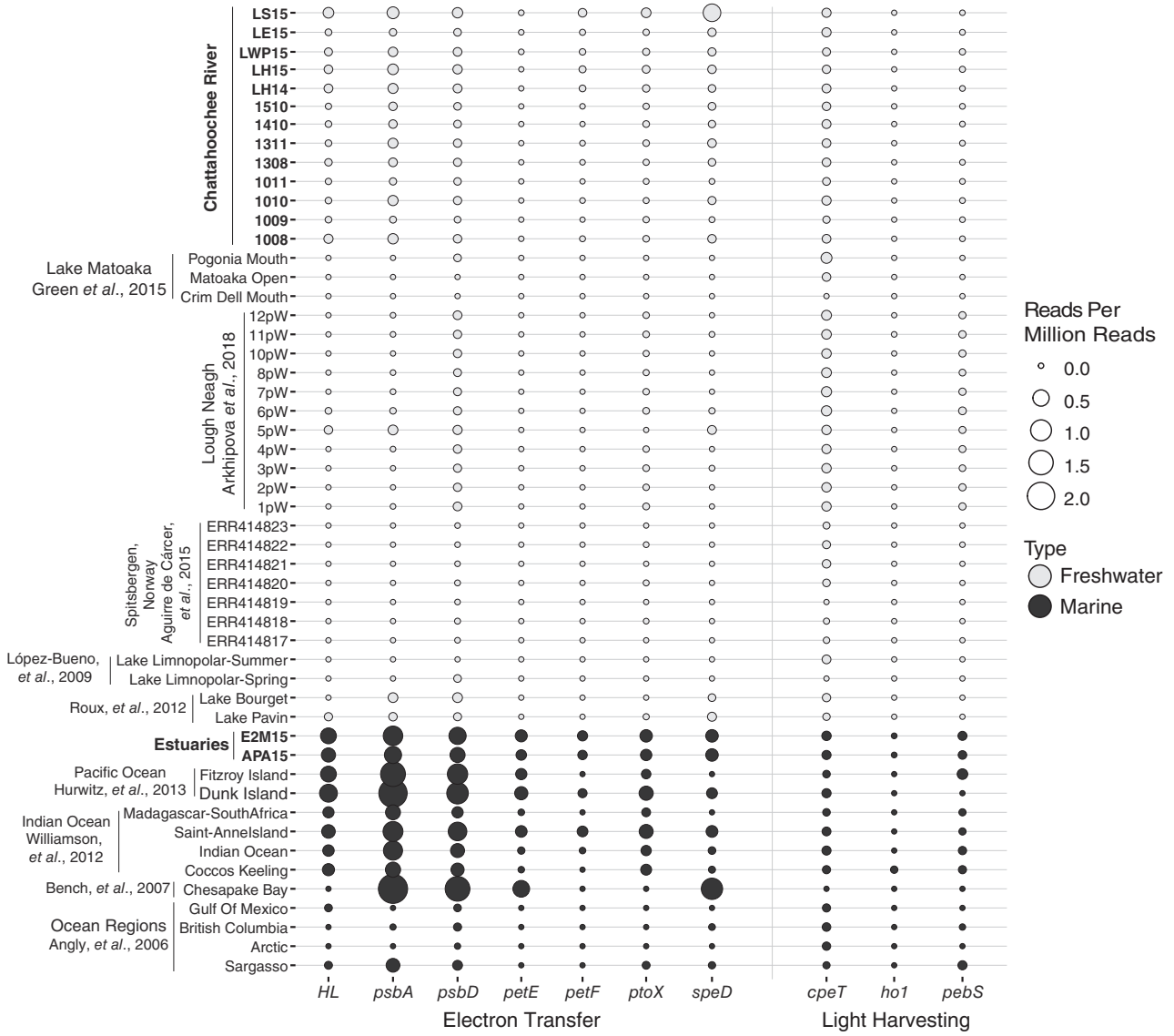

Fig. 3. Abundance profiles of photosynthesis-related AMGs in freshwater and marine environments.
Viral metagenomic reads were recruited to each AMG database using BlastX and filtered using ROCker position-specific bitscore thresholds. Read counts were normalized by gene size over data set size and scaled to reads per million reads in different freshwater (light bubbles) and marine (dark bubbles) environmental virome metagenomic datasets. The size of the bubbles indicates their abundance. Note the overall increase of AMG abundance in marine samples (including the estuary samples from this study) compared with freshwater environments.

Myoviridae genomes also had *pcyA*, *pebS*, *cpeT*, *ho1*, *petE*, *petF* and *ptoX* in addition to *psbA*, *psbD*, *hli* and *speD*, while most *Podoviridae* genomes only had *psbA* and *hli* (Supporting Information Table S1). Interestingly, *psbA* was also present in VC_1010_5 and VC_1011_18 (predicted to be circular and therefore probably a complete genome), which were classified as S-EIV1-like.

Considering the widespread detection of photosynthesis-related AMGs in viral genome fragments recovered along the Chattahoochee River, we assessed in greater detail their abundances using a read-based approach. Comparison of the relative abundances of AMG sequences in freshwater (Chattahoochee River, Lake Pavin and Bourget; Limnopolar lake; Arctic Spitsbergen, Norway; Lake Matoaka and Lough Neagh, Ireland), marine (Pacific Ocean, Indian Ocean, Arctic Ocean, British Columbia Marine Sample, Gulf of Mexico and Sargasso Sea) and estuarine virome datasets (Chesapeake Bay and estuaries at the Chattahoochee River delta) revealed an overall higher abundance of AMGs in marine environments than freshwater samples by approximately one order of magnitude (Fig. 3). This trend was also observed within the Chattahoochee River samples, with the estuarine locations Apalachicola Beach and East Point showing similar abundances of AMG genes to the marine datasets and higher abundances than the freshwater lakes along the river.

To assess the diversity of photosynthesis-related AMGs in our samples, the sequences for each AMG recovered here were clustered at 95% nucleotide identity level (Supporting Information **Table S3**). Out of the ten AMGs found in viral genomes along the Chattahoochee River, six recruited reads from most lake and estuarine locations (*petF*, *speD*, *ptoX*, *psbD*, *psbA* and *hli*) while the remaining only from the estuaries. Interestingly, all AMGs showed higher diversity, measured using the Shannon diversity index, towards the estuaries relative to the upstream freshwater lakes (Supporting Information **Fig. S4**). There was a clear relationship between fluvial distance and diversity in all AMGs except for *speD* (R^2 between 0.54 and 0.83, p-value <0.01). Although the estuaries have higher AMG diversity than the freshwater lakes, they are not solely responsible for the increase in diversity as demonstrated when the trend is maintained even when estuaries are removed from the analysis (p-value <0.01, Supporting Information **Fig. S4**). This suggested an accumulation of AMGs towards the estuaries that could be driven by the water flow along the basin or by the presence of cyanobacterial hosts in specific provinces along the river. The presence and abundance of AMGs recovered from freshwater lakes in the estuaries support the role of water flow as a driving factor for AMG diversity. However, a single case (*speD*) showed no evident relationship between fluvial distance and diversity towards the estuaries suggesting that water flow by itself might not be the only factor affecting AMG presence and maintenance along the river.

Evolutionary history of psbA *and related photosynthesis AMGs in freshwater viruses*

We further characterized the diversity of the identified freshwater viral *psbA* sequences by assessing their phylogenetic relationships with previously reported freshwater (clone libraries, PCR-derived sequences and virome surveys) and marine (PCR-derived sequences) viral and bacterial *psbA* sequences, in addition to reference genomes in the public databases or assembled from companion bacterial metagenomes from the Chattahoochee riverine system. Several of the reference and recovered sequences had significant signals of recombination consistent with what has been previously reported for *psbA* gene sequences (Chenard and Suttle, 2008) and bacterial and viral marine isolate genomes (Sullivan *et al*., 2006). These sequences were removed before constructing the *psbA* phylogenetic tree to isolate the effects of recombination. The phylogenetic reconstruction resulted in five distinct clades reflecting the genes' genomic context (i.e., bacterial or viral) and environmental origin (i.e., marine or freshwater) (Fig. 4 and Supporting Information **Fig**. **S5**). Overall, we found that there is a clear separation between bacterial (Clade I) and viral clades (Clades II, IV and V) except for Clade III that branched out from the marine *Synechococcus Myoviridae* cluster (Clade II) and was formed by *Prochlorococcus* and their infecting phages that were intermixed even at the deep branches.

Most freshwater-originating *psbA* sequences from the Chattahoochee ecosystem (20/30) clustered with sequences that have been recovered from lacustrine environments around the world in Clade IV (with only two *Myoviridae* isolate sequence exceptions; a marine, SRSM88 and a coastal, S-PM2), indicating an evolutionary separation between the marine and freshwater or estuarine versions of the gene. Cluster IV was represented by viral sequences from the Chattahoochee river lakes and only 4 (out of 40 in total) sequences from the estuaries, along with 16 sequences from the temperate Lake Erie (Midwest USA), 17 sequences from oligotrophic and oligomesotrophic temperate lakes Annecy & Bourget (France), Lake Ontario (Canada; n = 5), China paddy water (n = 13), Japan paddy water (n = 7) and one from Lough Neagh Lake (Ireland). This cluster also included a *psbA* sequence from *Synechococcus*-infecting viral isolates (MC15 and MC19) from Lake Erie and the only sequenced *psbA*-encoding *Synechococcus*-infecting freshwater viral isolate (S-CRM01) from the upper Klamath River (USA). Interestingly, the separation observed between the freshwater (Clade IV) and marine (Clade II) *psbA* sequences was restricted only to gene sequences originating from genomes classified as *Myoviridae,* while both marine and freshwater *psbA* sequences originating from *Podoviridae* genomes (assigned to previously reported Clade B based on major capsid protein analysis, Supporting Information **Fig. S6**) clustered together in Clade V. Clade V was represented by most estuarine sequences from the Chattahoochee River (Apalachicola Beach n = 12 and East Point n = 11); one sequence from Lake Seminole; several sequences from China (n = 18) and Japan (n = 8) rice paddy waters; sequences from the Red Sea, Mediterranean Sea, and Norway Coastal waters (n = 14); three brackish environment-isolated *Podoviridae* cyanophages S-CBP1, S-CBP3 and S-CBP4 (Chesapeake Bay, Maryland, USA) (Huang *et al*., 2015) and two coast-isolated *Podoviridae* cyanophages S-RIP1 and S-RIP2. Finally, one small cluster comprised of three almost identical sequences was found between the freshwater *Myoviridae* and *Podoviridae* clusters. One of the sequences in this cluster belonged to genome VC_1010_5, classified as S-EIV1-like. This group has not been previously reported to posses *psbA*. Not surprisingly, MAG-derived sequences clustered according to their bacterial origin in Clade I and the single case observed in which a MAG sequence clustered with phage sequences was resolved as phage origin using VIRSorter (a case of mis-assembly or binning).

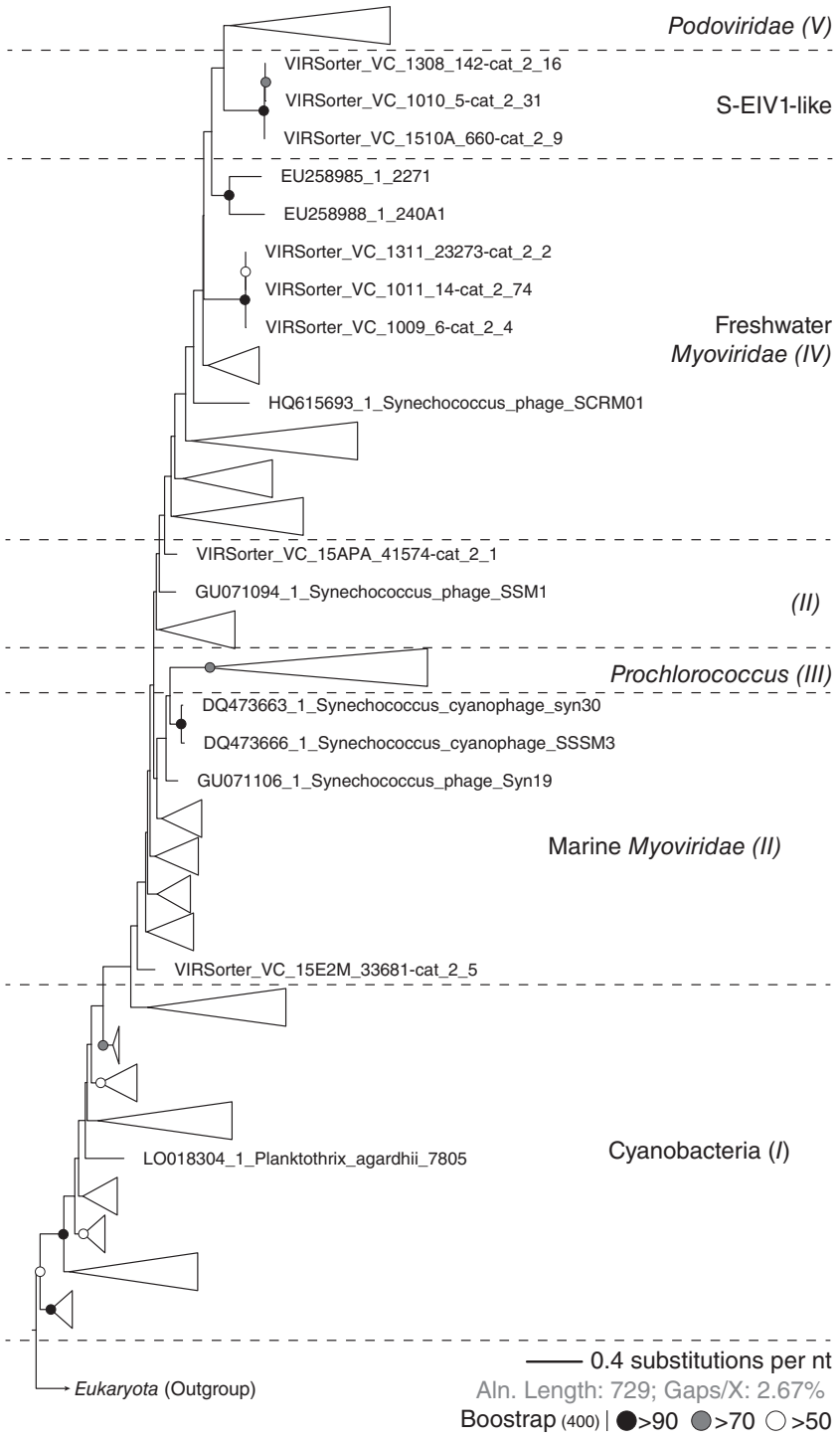

Fig. 4. Collapsed phylogenetic tree of *psbA* gene sequences.
Maximum likelihood phylogeny based on *psbA* gene nucleotide sequences of freshwater viral contigs from this study and previously determined freshwater- and marine-derived sequences. Major clustering groups are labelled on the right-hand side with the corresponding cluster number in parenthesis. Note the separation of freshwater *Myoviridae* (clade IV) and their marine counterparts (Clades II and III) as well as the separation of *Podoviridae* (clade V) and *Myoviridae* (clade II, III and IV) sequence clusters.

Inclusion of *psbA* sequences from the companion bacterial metagenome contigs into the phylogenetic tree (Supporting Information **Fig. S4**) revealed that 15 Lake Lanier sequences (including sequences retrieved from biomass collected on filters with pore size 1.6–2.5 μm) clustered together with sequences belonging to the Eukaryotic

fraction and to *Synechocystis*, *Microcystis*, *Nostoc*, *Anabaena*, and *Planktothrix*. This finding suggested that surveys of larger fractions of biomass are important in order to account for filamentous cyanobacteria that are not recovered when pre-filtration approaches are used. In general, the addition of sequences from the bacterial metagenome fraction did not change the overall tree topology with only a few changes in the branching pattern of ancestral branches with low support (unstable) being observed. That is, *Podoviridae* clustered in a single group (Clade V) next to the freshwater *Myoviridae* group (Cluster IV). Both groups were separated from both the marine *Myoviridae* (II) and the cyanobacterial (I) clusters; the latter being separated into two sub-clusters. Moreover, the intermediate S-EIV1-like group found between *Podoviridae* and *Myoviridae* recruited more sequences from the bacterial fraction suggesting that this group indeed possesses *psbA* and is not an artifact of the virome assembly, an interesting finding that merits future research. Finally, out of the 173 sequences associated with any viral cluster, 5 were classified as viral by VIRSorter while the remaining 168 were not presumably due to their length being too short for robust classification (≤ 5000 bp, n = 135). The remaining 33 sequences longer than 5000 bp not classified as viral by VIRSorter had viral sequences as their best Blastp matches against the NCBI's nr database indicating that all of them were likely of viral origin.

The remaining photosynthesis-related AMGs showed similar trends of separation between bacterial and viral versions with the exception of SpeD (Supporting Information **Figs. S6–S8**). Although there was clear separation between viral and bacterial versions, only the *psbA* gene sequence phylogeny further supported the separation between the marine and freshwater versions. Moreover, considering that *psbA* also allowed us to taxonomically robustly discriminate between *Myoviridae* and *Podoviridae* gene sequences (Fig. 4 and Supporting Information Fig. **S5**), *psbA* clearly showed the best discriminatory power for tracking purposes compared with the rest of photosynthesis-related AMGs.

Discussion

Large unexplored cyanophage diversity shows endemism along the Chattahoochee River

Viral genome fragments recovered along the Chattahoochee River classified using MetaVir and vConTACT2 showed that most of the photosynthesis related AMGs were encoded by complete or partial *Myoviridae* and *Podoviridae* genomes that represented novel viral genera and species based on the recovery of 4 VCs with no representatives in the public databases. Genome classification showed that *Myoviridae* fragments originated from all samples suggesting that *Myoviridae* cyanophages are present along the river basin while *Podoviridae* cyanophages, originating mostly from the estuaries, were not common in freshwater lake locations. Interestingly, we observed that two of the Chattahoochee River genomes clustered with a recently described phage belonging to a novel family (S-EIV1) (Chenard *et al*., 2015) and encoded *speD* and *psbA*, which has not been reported for this novel lineage previously. These results likely reveal a recent lateral transfer event of these genes between cyanophage lineages. On the other hand, multiple criteria used to predict phage–host interactions resulted in ~2% of the phage genomes recovered being assigned to a potential bacterial host (Supporting Information **Table S4**). Given the variable success of these approaches (Edwards *et al*., 2016; Paez-Espino *et al*., 2016; Emerson *et al*., 2018), the low frequency of host assignment was not surprising. These findings highlighted that the identification of phage–host associations based on metagenomic data remains challenging and the need for further improvement in this area of research.

In general, read-based genome abundance analysis showed that most AMG-containing genome fragments appear to be endemic to a specific region along the river (83.54%), with few exceptions that showed no endemicity, i.e., they were rare or present in all locations. The dispersion limitation for most of the genomes was not strong in the lakes downstream of Lake Lanier, suggesting increased intermixing of phages in downstream lakes, probably influenced by the relatively short distance between nearby dams. In contrast to these downstream lakes, a distinct viral province in the most upstream Lake Lanier was observed, supported by the difference in genomes endemic to Lanier relative to the other freshwater sites. Consistent with these interpretations, we have observed similar biogeographic patterns along the river basin and the existence of a Lake Lanier microbial province in the companion bacterial metagenomes (i.e,. the potential viral hosts) recovered from the same sampling sites (Tsementzi *et al*., in press). Overall, our results demonstrated that viral diversity closely followed that of the accompanying bacterial community (Supporting Information **Fig. S10**), and both tended to increase towards the estuaries suggesting an effect of river flow in the accumulation of both bacterial and viral species in downstream locations.

The factors that limit microbial dispersion and therefore, the structure of viral communities from Lake Lanier currently remain elusive but could be related to the anthropogenic effects of the Atlanta Metro area, the largest municipality in the region located just downstream of Lake Lanier (and upstream of the remaining four lakes), on water (bio-)chemistry. In addition, the hydrology of this riverine ecosystem could play an important role in

shaping the microbial community patterns observed (Nino-Garcia *et al*., 2016). For instance, each lake is formed by a manmade dam that limits the flow of water for hydroelectric energy production. It is important to note that when some dam doors are opened, the water that flows downstream drains from the bottom of the lake (all our samples were surface water). Further, the amount of water that flows at regular intervals varies between ~3,785 and ~37,850 l/s in several of the dams and can reach flow rates up to 300,000 l/s in Lake Seminole dam (during the years sampled, U.S. Geological Survey, www.usgs.gov), which could also have an effect on the dispersal of microbial and viral populations. Therefore, even if the viral populations are indeed endemic in their province, it is plausible to find them in any given sampling period considering their high particle densities *in situ* (10^6–10^8 viruses per millilitre of water) (Zhong *et al*., 2014; Mohiuddin and Schellhorn, 2015).

Photosynthesis AMGs are widely distributed and prevalent in freshwater environments

Viral AMGs, especially *psbA* homologues, were first reported in marine environments (Mann *et al*., 2003). Since then, they have been recognized as an important mechanism for increased particle production and host evolution (Lindell *et al*., 2004). The presence of these photosynthetic genes in viral genomes from the Chattahoochee River suggested that viral particle production strategies used by freshwater cyanophages might be similar to marine cyanophages. The high prevalence of *psbA*, *psbD*, *speD* and *hli* genes in freshwater or marine viral communities confirms that cyanophages benefit from possessing genes that provide a steady supply of transcripts and proteins for the assembly of photosystem II machinery during infection (*psbA* and *psbD*) (Lindell *et al*., 2005). The viral PsbA version commonly found in cyanophages (similar to the stress-induced host D1:2 isoform) has also been hypothesized to be less susceptible to turnover and photodamage (Clarke *et al*., 1993; Sullivan *et al*., 2006), providing the bacterial host with the necessary proteins to continue photosynthesis even during infection (Sharon *et al*., 2007). Moreover, the fact that *speD* and *hli* genes were present in a high proportion in the freshwater viral genomes recovered here indicated that protection of the photosystem machinery against high-light excitation damage by *hli* proteins (He *et al*., 2001; Lindell *et al*., 2004; Komenda and Sobotka, 2016), or maintenance of the machinery's activity, structure and adaptation by polyamines synthesized by *speD* (Bograh *et al*., 1997; Gao *et al*., 2016) is an additional important feature likely under positive selection to increase viral particle production in the dying host as in marine cyanophages. However, the high prevalence of *speD* in viral genomes, which clustered with other non-cyanosiphoviruses in the vConTACT2 classification (data not shown) suggested that *speD*, involved in the production of spermidine, might not be restricted to its possible role in photosynthesis as previously speculated (Wortham *et al*., 2007). Nonetheless, removal of the viral genomes encoding only SpeD did not affect our major conclusions (but substantially reduced the number of viral genomes analysed).

Myoviridae genomes encoded most of the photosynthesis AMGs surveyed in this study while *Podoviridae* genomes encoded only *psbA* and *hli*. Due to the fragmented nature of the recovered genomes, it is possible that *Podoviridae* in freshwater environments encode a wider range of AMGs not recovered in the assembly data of our study. However, three circular genomes recovered here and assigned to *Podoviridae* by Metavir and vConTACT also encoded a reduced set of photosynthesis AMGs. Further, these findings were consistent with previous reports showing that members of the *Podoviridae* family only encode a reduced set of photosynthesis AMGs (Puxty *et al*., 2015). This characteristic is probably due to shorter latent periods of *Podoviridae* that are likely too short for complete production and activation of the photosynthesis genes compared to *Myoviridae* cyanophages (Mann, 2003; Sullivan *et al*., 2006). Consistent with this interpretation, the lower abundance of freshwater AMGs compared with the marine counterparts may reflect the higher nutrient load (and shorter latent periods) of riverine environments than the ocean (Bristow *et al*., 2017). Indeed, Chattahoochee River samples showed higher nutrient content (~0.07 mg/l NO_2 + NO_3 and ~0.13 mg/l PO_4, [Supporting Information Table S6]) than those typically observed in the Gulf of Mexico surface waters (below detection for NO_2 + NO_3 and ~0.014 mg/l PO_4) (Tsementzi *et al*., 2016). This would allow increased growth rates for bacterial hosts and, therefore, select for shorter latent periods of cyanophages as previously reported for a *Podoviridae* member (Sullivan *et al*., 2006). However, this hypothesis remains to be experimentally tested in order to determine if the latent period is indeed different between these novel, freshwater cyanophage species and genera and their marine counterparts.

The fact that not all photosynthesis-related AMGs from ocean viruses (Puxty *et al*., 2015; Crummett *et al*., 2016; Gao *et al*., 2016) were recovered in our freshwater viromes (i.e., *pcyA*, *psaA*, *ho1*, and *nblA* genes) suggested that there might be different ecological strategies and/or physiologies of freshwater cyanobacteria and their viruses, compared with their marine counterparts, which warrant further investigation. Phage–host prediction and linkage is an important first step in understanding the presence and abundance of AMGs in phages. The complete and partial viral genome sequences reported here should enable future studies on the infection strategies and host

preferences (e.g., unicellular vs. filamentous) used by freshwater cyanophages in understudied freshwater ecosystems.

Viral AMGs alleles differentiated from bacterial versions may serve as viral biomarkers

In general, viral photosynthesis-related AMGs, especially the *psbA* gene, appear to be clearly separated from their bacterial homologues. In agreement with these results, previous studies in marine isolates showed a clear separation between phage and host sequence clusters (Sullivan *et al*., 2006) suggesting that, based on the sequences studied here, there are infrequent recent HGT events of AMGs between hosts and viruses, and that indeed the viral versions of these genes have evolved and been maintained for substantial evolutionary time in the phage genetic pool (Sandaa *et al*., 2008). An exception to this pattern was observed in the marine *Prochlorococcus psbA* cluster (Cluster III), which contained both phage and bacterial sequences branching from the marine *Synechococcus Myoviridae* cluster (Cluster II) revealing the possibility of a more recent transfer event between phages and hosts. Consistent with our expectations, the inclusion of *psbA* gene sequences from the bacterial fraction showed that no bacterial sequence clustered along with viral-derived sequences (with the caveat of sequences from short contigs that could not be robustly classified as viral nor confirmed as bacterial). It is important to note, however, that the evolutionary relationships between *psbA* gene sequences and phages encoding them studied here were also complicated by the presence of introns and homing endonucleases next to or within *psbA* sequences often having an algl origin (Millard *et al*., 2010), which warrants future investigations because it implies possible interactions (ancient or recent) with algal populations. We excluded these sequences from the final phylogeny (e.g., Fig. 4 and Supporting Information **Fig**. **S5**) to avoid additional complications.

Furthermore, our analysis showed that our genomes shared a significant portion of their content with marine *Myoviridae* and *Podoviridae* phages and that a few freshwater *psbA* sequences clustered within marine clades highlighting the possibility that cyanophages may be moving between these two distinct habitats, for instance, as part of the river flowing to the ocean. Notably, previous experiments have shown that freshwater viruses can be propagated in marine organisms (Sano *et al*., 2004), consistent with the findings reported here and highlighting the adaptability of some of the cyanophages recovered here. However, the predominant signal from the phylogenetic analysis performed here showed that viral *psbA* genes can cross, but not thrive in different habitats (e.g., presence of freshwater sequences in estuarine samples), since most *psbA* gene sequences still clustered based on their origin (e.g., distinct freshwater and ocean sequence clusters) and the genomes that encoded them showed strong endemicity. These clustering patterns along with the genome abundance patterns highlighted earlier suggested that cyanophages, for the most part, tend to be widespread along the river but their abundance and prevalence are shaped and maintained by the environmental selection pressures acting on them and/or their host within distinct provinces. Our findings are thus consistent with the tenet that microbes and their accompanying phages can be found everywhere; however, their relative abundance and therefore our ability to detect them depends on the environment.

Previous studies have shown that cyanobacteria probably originated in freshwater environments (Blank and Sanchez-Baracaldo, 2010). This raises the question: where did the first viral *psbA* genes appear, i.e., in freshwater or oceanic systems? Sequence analysis of the only *psbA*-encoding freshwater phage isolate genome (S-CRM01) against its marine counterparts indicated that the acquisition of *psbA* by freshwater phages was not a recent event (Dreher *et al*., 2011). Accordingly, the phylogenetic analysis of *psbA* genes (when present) could provide robust means for assessing the habitat of origin of phages. For instance, phage MC15 isolated from Lake Erie (Wilhelm *et al*., 2006) appeared to be more closely related to marine phages based on previous g20 viral capsid protein phylogenies. However, it was placed in the freshwater clade based on our *psbA* gene phylogeny, which confirmed its freshwater origin. The genomic and genetic information presented in this study could begin to shed light onto the origin of viral photosynthesis gene transfers from host to phages and from/to different environments.

Overall, our results demonstrated the presence and high diversity of viral *psbA* genes in freshwater ecosystems as well as the previously underestimated presence of these genes in coastal/estuarine representatives of the *Podoviridae* family. The ecological importance of *psbA*, revealed by its prevalence in marine and freshwater habitats around the world, suggested fitness advantages for phages carrying the gene in freshwater lakes, estuaries and oceans. The estuarine- and freshwater-enriched OTUs indicated that habitat-specific adaptation is ongoing, and explained to a large extent, the distribution patterns of the OTUs recovered. Collectively, our findings better delineate the *psbA* sequence types that are predominant in different water bodies and substantially expand the known sequence diversity of *psbA* genes.

While the relatively lower abundance of *psbA* genes in freshwater versus estuarine and oceanic ecosystems could be explained by the nutrient-rich nature of the habitat and/or different ecological strategies, alternative explanations related to host–phage interactions or host population dynamics cannot be discarded. Therefore,

whether the signal of photosynthesis AMGs, in particular of the *psbA* gene, was associated with cyanophages infecting only unicellular freshwater populations of *Synechococcus* or also reflected other phages infecting filamentous bacterial species remains to be investigated. Future studies should focus on the host–phage infection networks in freshwater ecosystems to ecologically contextualize these findings. The genome and gene sequences reported here should greatly facilitate such studies, e.g., by providing sequences for qPCR assay design and monitoring.

Experimental procedures

Sample collection and DNA extraction

Viral metagenomic samples (or 'viromes') were obtained from seven sites across the approximately 760 km-long Chattahoochee River from 2010 to 2015 (Fig. 1). All samples were collected in the late summertime (September–October). In addition, August and November samples from the most upstream lake, Lake Lanier (GA), were also included for comparison. Estuarine samples were taken at approximately 38 km upstream of the open ocean. For each sample, 10 l of water were collected in acid-washed carboys, transferred to the laboratory and stored at 4°C until water filtration was performed within 24 h from collection time. Water was pre-filtered through sterivex filters (0.2 μm porosity) with a peristaltic pump, and the filtrate was subsequently processed to concentrate viral particles as previously described (John *et al*., 2011). Briefly, 1 ml ferric chloride solution (10 g Fe/l) was added to each 10 l of water and samples were mixed for a minimum of 2 h on a stir plate. The resulting viral flocculate was isolated on a 142 mm 0.8 μm porosity filter (Pall Supor-800 filter), and filters were stored in AEM buffer (0.1 M EDTA; 0.2 M MgCl2; 0.2 M Ascorbate, pH 6.0) at 4°C until further processing. Typically, 10 ml of viral concentrate was obtained for each sample (from the initial 10 l of lake water). Concentrated viral particles were subsequently treated with DNase to eliminate free bacterial DNA, and further purified in CsCl step gradients as previously described (Hurwitz *et al*., 2013). Step gradients were built in 38.5 ml capacity ultracentrifuge tubes (Ultra-Clear™ tubes, Beckman Coulter), and centrifuged for 4 h at 24,000 rpm in a Beckman Coulter Optima L-90K Ultracentrifuge (SW 32 Ti Rotor). The resulting phage bands were collected from the interface of the 1.4 and 1.56 g/ml step gradients until no visible particles were observed. Subsequently, additional 1 ml aliquots were drawn from the interface and added to the previously collected bands if their density was within the desired range (1.4–1.56 g/ml). The collected bands were purified with step dialysis as described previously (Hurwitz *et al*., 2013). Typically, 10 ml of purified and dialyzed viral particles for each sample was obtained and further concentrated to ~3 ml using Amicon 100kDa nominal molecular weight cutoff filters. DNA was extracted using the Omega Viral MagBind Kit following the manufacturer's instructions.

Sequencing, assembly and processing

DNA sequencing libraries were prepared using Illumina's Nextera XT DNA library prep kit per the instructions of the manufacturer, except that the protocol was terminated after isolation of cleaned amplified double-stranded libraries. Library concentrations were determined using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific) and average insert sizes were determined using the Bioanalyzer 2100 instrument (Agilent). Libraries were sequenced on an Illumina HiSEQ 2500 instrument (2 x 150 bp paired-end) using the HiSeq Rapid PE Cluster Kit v2 and HiSeq Rapid SBS Kit v2 (Illumina). Adapter trimming and demultiplexing of samples was carried out by the instrument.

Resulting sequences were quality filtered using trimmomatic (Bolger *et al*., 2014) and in-house scripts of the enveomics collection (Rodriguez-R and Konstantinos, 2016) to remove low-quality sequences, which were defined as reads with an average quality score below 25 and length < 50 bp. High-quality sequences were assembled using IDBA-UD v.1.1.1 with default parameters (kmer size 20–100) (Peng *et al*., 2012). Bacterial contamination measured as the percentage of 16S rRNA gene-encoding reads was estimated using ParallelMeta v.3.4.3 (Jing *et al*., 2017). Virome datasets were considered not contaminated when they contained less than 0.02% of 16S rDNA reads (Roux *et al*., 2013; Enault *et al*., 2017). Resulting contigs larger than 500 bp were analysed using VIRSorter (Roux *et al*., 2015) with the virome decontamination mode and Metavir (Roux *et al*., 2014) for classification and annotation. Briefly, VIRSorter identifies phage contigs based on two metrics: the first being the presence of hallmark viral genes or enrichment of viral or non-*Caudovirales* genes and the second metric includes characteristics of the contigs such as depletion in PFAM affiliated genes, enrichment of short and/or uncharacterized genes and depletion in strand switch. Based on these metrics, VIRSorter assigns contigs in three categories: Category 1, most confident viral; Category 2, 'likely viral' predictions; Category 3, 'possible viral' predictions.

Identification of photosynthesis-related genes

An *in-house* database for each AMG encoded protein involved in photosynthesis (photosystem-related – PsbA, PsbD, PsaA, SpeD and high-light inducible proteins; electron transport – PTOX, PetE and PetF; photosynthetic pigments – heme oxygenase 1, PebS, PcyA, CpeT

and NblA) was constructed using protein sequences from marine and freshwater cyanobacterial genomes and their infecting viruses retrieved from UniProt. Genes from the genome fragments assembled as part of this study predicted with Prodigal v.2.6.1 (Hyatt *et al*., 2012) were searched against the *in-house* database using Blastp (Altschul *et al*., 1990). The best match of each gene was considered when the match had Bitscore >=70, ID >=40%, and coverage of target protein >=50%, thresholds previously shown to minimize false positives (Enault *et al*., 2017). The sequences were also manually verified by visually inspecting protein and DNA alignments before being selected for further analysis. Only contigs encoding a viral photosynthesis AMG sequence (based on Blastp results and manual analysis) and assigned to category 1 or 2 by VIRSorter were used for further analysis. These *in-house* databases also included photosynthesis AMGs derived from companion bacterial metagenome-assembled genomes (MAGs) previously recovered from the same Chattahoochee River samples as the viromes (Tsementzi *et al*., 2019) as well as AMGs recovered in the assembly (but not binned into MAGs) of these metagenomes. The cutoff for a match used in the latter case was of higher identity threshold (≥ 70% identity) for increased stringency in identifying AMGs of truly bacterial origin.

Classification of viral contigs encoding photosynthesis-related AMGs

Contigs containing photosynthesis-related AMGs were de-replicated at 95% average nucleotide identity threshold, on 80% or more on the length of the shortest fragment, in order to reduce the number of redundant viral genomes using CD-HIT (Li and Godzik, 2006; Fu *et al*., 2012). After de-replication, short genome fragments (<5Kbp) were removed from subsequent analysis. Genome classification was performed using vConTACT2 (Bolduc *et al*., 2017; Jang *et al*., 2019) and the ViralRefSeq-prokaryotes-v85 database containing 2305 prokaryotic phages. Briefly, vConTACT2 uses a Markov clustering approach on the proteins shared between viral genomes resulting from an all vs all Blastp search (e-value ≥1E-5 and Bitscore ≥50) and represents the genomes in a weighted network. Genome clusters obtained by this approach are considered a good proxy for genus-level grouping of related viral genomes. The resulting networks from the vConTACT2 classification were visualized with Cytoscape (version 3.7.0; http://cytoscape.org/) as previously reported (Bolduc *et al*., 2017).

To assess the degree of endemism of AMG-encoding genomes, the abundance of viral contigs in different locations and sampling years was estimated based on the number of recruited reads. Briefly, sequencing reads from each viral metagenome were mapped against AMG-encoding genome fragments using BLAT (Kent, 2002) with the -fastMap flag and considering only the best match for each read. Genome abundances were estimated as sequencing read counts mapped to each genome and normalized by total-sum scaling, i.e., the total number of reads per virome. Subsequently, cubic smoothing splines were estimated based on the log-normalized abundances for each genome and sampling site/lake to account for sample heterogeneity and stochasticity between sampling years. Finally, the abundance splines were regressed against the location of the sampling site (i.e., distance); the locations were organized based on their distances from Lake Lanier, the most upstream site sampled. Genomes with Pearson's correlation index between the spline-derived (fitted curve to both abundance points) and the observed abundances at each site below 0.5 were discarded as non-endemic, e.g., transient, rare or cosmopolitan.

AMG-encoding viruses were assigned to a (putative) host based on a combination of criteria as previously reported (Paez-Espino *et al*., 2016; Roux *et al*., 2016; Emerson *et al*., 2018). Briefly, CRISPR loci were first identified in MAGs from the Chattahoochee River using the metagenome version of the CRISPR Recognition Tool (Bland *et al*., 2007; Rho *et al*., 2012), and subsequently searched against the viral genomes for a match using Blastn and a threshold of 100% nucleotide identity along the whole CRISPR spacer. Second, tRNAs present in the viral genomes were identified using Aragorn v.1.2.38 (Laslett and Canback, 2004), and subsequently searched against MAGs for a 100% nucleotide identity match. Hosts were identified when one of the above criteria was tested positive.

Abundance and diversity of AMGs in viral metagenomes

Quantification of photosynthesis-related AMGs in viral metagenomes from different freshwater and marine environments was performed using ROCker profiles (Orellana *et al*., 2017) built based on UniProt-derived AMG sequences. The ROCker-generated model calculates the most discriminant position-specific bitscore threshold in the protein alignment by simulating *in silico* metagenomes with a predefined read length. This strategy allows the accurate estimation of the abundance of target genes in short read data while decreasing the number of false positives. Protein sequences encoded on the reads were predicted by FragGeneScan (Rho *et al*., 2010). AMG abundances were compared with other marine and freshwater metagenomic data sets (Angly *et al*., 2006; Bench *et al*., 2007; Lopez-Bueno *et al*., 2009; Roux *et al*., 2012; Williamson *et al*., 2012; Hurwitz and Sullivan, 2013;

Aguirre de Carcer *et al*., 2015; Green *et al*., 2015; Arkhipova *et al*., 2018), following the same approach.

Alpha diversity analyses were performed for each AMG by clustering sequences at 95% nucleotide identity using VSEARCH v.2.9.0 (Rognes *et al*., 2016). Reads previously identified by ROCker were mapped onto gene clusters to assess the richness (observed and estimated using Chao1) and diversity (Shannon index) in our samples (Chao *et al*., 2010). OTU abundance (reads mapped) was normalized by OTU length and dataset size, and expressed as reads per 1Kbp per million reads. Regression analyses were performed between fluvial distance along the Chattahoochee River and diversity estimates of AMG genes using R (v.3.4.4). The analysis was repeated in rarefied samples accounting for dataset size differences. Viral and bacterial metagenome sequence diversity was calculated using Nonpareil v.3, which estimates the coverage of a metagenome based on the level of read redundancy, and derivates a measure of alpha diversity that has been shown to correlate well with 16S rRNA gene diversity profiles (Rodriguez *et al*., 2018).

Phylogenetic reconstruction of psbA *and photosynthesis-related AMG sequences*

For *psbA*, cyanobacterial host (n = 99) and viral (n = 64) reference sequences obtained from the NCBI GenBank database were supplemented with sequences from previous studies and sequences derived from high-quality (bacterial) MAGs from the Chattahoochee River (Tsementzi *et al*., 2019), for a total set of 518 bacterial and viral *psbA* reference sequences from marine and freshwater environments. This database included viral sequences from freshwater environments: paddy water incubation experiments (PCR amplification – cloning), China (Wang *et al*., 2016), rice floodwater (PCR amplification, D-DGGE), Japan (Wang *et al*., 2009), Lake Annecy and Bourget (PCR amplification, DGGE), France (Zhong and Jacquet, 2013), Lake Constance, Germany; Experimental Lakes Area, Ontario, Canada (PCR amplification) (Chenard and Suttle, 2008), Lake Erie (PCR amplification), USA (Wilhelm and Matteson, 2008), Lake Lough Neagh (Virome), Northern Ireland (Arkhipova *et al*., 2018), Lake Kinneret (PCR amplification), Israel (Junier *et al*., 2007) and the new sequences from the Chattahoochee River identified by our study, as well from marine environments, Hawaii Ocean Time Series (HOT, PCR amplification) (Sullivan *et al*., 2006), Norwegian coastal waters (PCR amplification) (Sandaa *et al*., 2008), Mediterranean and Red Sea (gene cloning) (Zeidner *et al*., 2005; Sharon *et al*., 2007). Data set sequences were aligned using MAFFT v.7.245 (Katoh and Standley, 2013) and tested for recombination signal using SplitsTree v.4.14.4 (Huson and Bryant, 2006). For data sets with positive recombination signal by SplitsTree analysis, recombinant sequences were identified, as suggested previously (Sullivan *et al*., 2006), using Recombination Detection Program (RDP) v.4.80 (Martin *et al*., 2015) as those for which at least two out of four methods implemented in RDP, i.e., GENECONV (Padidam *et al*., 1999), MaxChi (Smith, 1992), Chimaera (Posada and Crandall, 2001) and RDP (Martin and Rybicki, 2000), resulted in significant values for recombination after multiple testing correction. Such recombinant sequences were removed from building the final reference phylogenetic tree of freshwater and marine *psbA* sequences to avoid effects on the tree topology (294 sequences removed in total). In addition, sequences that were shorter than 500 bp were excluded from the phylogenetic reconstruction analyses. In several genomes, the recovered *psbA* sequences were split in two (or more) different segments. These gene sequences were also excluded from the phylogenetic tree to avoid any interference from possible pseudogenes that could affect the phylogenetic signal.

Maximum likelihood phylogenetic trees were constructed using RAxML v.8.2.12 (Stamatakis, 2014), removing identical sequences as recommended. A general time reversible model was used for *psbA* and *psbD* phylogenetic trees assuming gamma distributed variation rates among sites and proportion of invariable sites estimated by RAxML along with the -autoMRE option to determine the best number of bootstrap repetitions. The phylogenetic trees for the remaining AMGs were constructed in a similar way except for the use of protein sequences and the -PROTGAMMAAUTO parameter allowing RAxML to calculate the best substitution model for each data set. Phylogenetic trees with added AMG sequences from the bacterial fraction were also calculated using RAxML with the same parameters.

Data availability

In-house protein sequence databases used for Blastp searches and recovered AMG sequences are available at http://enve-omics.ce.gatech.edu/data/viralAMG. High-quality MAGs used in this study (Supporting Information Table S2), distances, and other taxonomic analyses are available at http://microbial-genomes.org/projects/WB_binsHQ. All metagenomic data sets from the Chattahoochee riverine ecosystem that were used in this study (Fig. 1) are available in the NCBI SRA database as part of the BioProjects PRJNA497294 (BioSamples SAMN10261643, SAMN10261646 to SAMN10261653, SAMN10261657 to SAMN10261659, SAMN10261661, SAMN10261664 to SAMN10261669, SAMN10261672, SAMN10261673).

Acknowledgements

Our work was supported, in part, by US NSF (awards 1416673 and 1759831 to KTK) and the Gordon and Betty

Moore Foundation (award 3790 to MBS). Special thanks to Ben Bolduc for his support and help with vConTACT2.

References

Aguirre de Carcer, D., Lopez-Bueno, A., Pearce, D.A., and Alcami, A. (2015) Biodiversity and distribution of polar freshwater DNA viruses. *Sci Adv* **1**: e1400127.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

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

Arkhipova, K., Skvortsov, T., Quinn, J.P., McGrath, J.W., Allen, C.C., Dutilh, B.E., McElarney Y., Kulakov L.A. (2018) Temporal dynamics of uncultured viruses: a new dimension in viral diversity. *ISME J* **12**: 199–211.

Bench, S.R., Hanson, T.E., Williamson, K.E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K.E. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. *Appl Environ Microbiol* **73**: 7629–7641.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**: 209.

Blank, C.E., and Sanchez-Baracaldo, P. (2010) Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. *Geobiology* **8**: 1–23.

Bograh, A., Gingras, Y., Tajmir-Riahi, H.A., and Carpentier, R. (1997) The effects of spermine and spermidine on the structure of photosystem II proteins in relation to inhibition of electron transport. *FEBS Lett* **402**: 41–44.

Bolduc, B., Jang, H.B., Doulcier, G., You, Z.Q., Roux, S., and Sullivan, M.B. (2017) vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**: e3243.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bragg, J.G., and Chisholm, S.W. (2008) Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. *PLoS One* **3**: e3550.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018) Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**: 754–766.

Bristow, L.A., Mohr, W., Ahmerkamp, S., and Kuypers, M.M. M. (2017) Nutrients that limit growth in the ocean. *Curr Biol* **27**: R474–R478.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., *et al*. (2015) Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498.

Chao, A., Chiu, C.H., and Jost, L. (2010) Phylogenetic diversity measures based on Hill numbers. *Philos Trans R Soc Lond B Biol Sci* **365**: 3599–3609.

Chenard, C., and Suttle, C.A. (2008) Phylogenetic diversity of sequences of cyanophage photosynthetic gene *psbA* in marine and freshwaters. *Appl Environ Microbiol* **74**: 5317–5324.

Chenard, C., Chan, A.M., Vincent, W.F., and Suttle, C.A. (2015) Polar freshwater cyanophage S-EIV1 represents a new widespread evolutionary lineage of phages. *ISME J* **9**: 2046–2058.

Clarke, A.K., Soitamo, A., Gustafsson, P., and Oquist, G. (1993) Rapid interchange between two distinct forms of cyanobacterial photosystem II reaction-center protein D1 in response to photoinhibition. *Proc Natl Acad Sci U S A* **90**: 9973–9977.

Crummett, L.T., Puxty, R.J., Weihe, C., Marston, M.F., and Martiny, J.B. (2016) The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology* **499**: 219–229.

Deng, L., and Hayes, P.K. (2008) Evidence for cyanophages active against bloom-forming freshwater cyanobacteria. *Freshwater Biol* **53**: 1240–1252.

Dreher, T.W., Brown, N., Bozarth, C.S., Schwartz, A.D., Riscoe, E., Thrash, C., Bennett S.E., Tzeng S.C., Maier C. S. (2011) A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. *Environ Microbiol* **13**: 1858–1874.

Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016) Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev* **40**: 258–272.

Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B. J., Jang, H.B., Singleton C.M., Solden L.M., Naas A.E., Boyd J.A., Hodgkins S.B., Wilson R.M., Trubl G., Li C., Frolking S., Pope P.B., Wrighton K.C., Crill P.M., Chanton J. P., Saleska S.R., Tyson G.W., Rich V.I., Sullivan M.B. (2018) Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* **3**: 870–880.

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M.B., and Petit, M.A. (2017) Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* **11**: 237–247.

Fernandez, L., Rodriguez, A., and Garcia, P. (2018) Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J* **12**: 1171–1179.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.

Gao, E.B., Gui, J.F., and Zhang, Q.Y. (2012) A novel cyanophage with a cyanobacterial nonbleaching protein A gene in the genome. *J Virol* **86**: 236–245.

Gao, E.B., Huang, Y., and Ning, D. (2016) Metabolic genes within Cyanophage genomes: implications for diversity and evolution. *Genes (Basel)* **7**, 80.

Green, J.C., Rahman, F., Saxton, M.A., and Williamson, K. E. (2015) Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA. *Aquat Microb Ecol* **75**: 117–128.

Gregory, A.C., Solonenko, S.A., Ignacio-Espinoza, J.C., LaButti, K., Copeland, A., Sudek, S., Maitland A., Chittick L., dos Santos F., Weitz J.S., Worden A.Z., Woyke T., Sullivan M.B. (2016) Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* **17**: 930.

He, Q., Dolganov, N., Bjorkman, O., and Grossman, A.R. (2001) The high light-inducible polypeptides in *Synechocystis* PCC6803. Expression and function in high light *J Biol Chem* **276**: 306–314.

Hellweger, F.L. (2009) Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. *Environ Microbiol* **11**: 1386–1394.

Huang, S., Zhang, S., Jiao, N., and Chen, F. (2015) Comparative genomic and phylogenomic analyses reveal a conserved core genome shared by estuarine and oceanic Cyanopodoviruses. *PLoS One* **10**: e0142962.

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

Hurwitz, B.L., and U'Ren, J.M. (2016) Viral metabolic reprogramming in marine ecosystems. *Curr Opin Microbiol* **31**: 161–168.

Hurwitz, B.L., Deng, L., Poulos, B.T., and Sullivan, M.B. (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol* **15**: 1428–1440.

Huson, D.H., and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254–267.

Hyatt, D., LoCascio, P.F., Hauser, L.J., and Uberbacher, E. C. (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**: 2223–2230.

Igarashi, K., and Kashiwagi, K. (2010) Modulation of cellular function by polyamines. *Int J Biochem Cell Biol* **42**: 39–51.

Jang H.B., Bolduc B., Zablocki O., Kuhn J., Roux S., Adrianenssens, E., *et al*. (2019). Gene sharing networks to automate genome-based prokaryotic viral taxonomy. bioRxiv:533240.

Jing, G., Sun, Z., Wang, H., Gong, Y., Huang, S., Ning, K., Xu J., Su X. (2017) Parallel-META 3: comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. *Sci Rep* **7**: 40371.

John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K., Kern, S. et al. (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* **3**: 195–202.

Junier, P., Witzel, K., and Hadas, O. (2007) Genetic diversity of cyanobacterial communities in Lake Kinneret (Israel) using 16S rRNA gene, *psbA* and *ntcA* sequence analyses. *Aquat Microb Ecol* **49**: 233–241.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656–664.

Komenda, J., and Sobotka, R. (2016) Cyanobacterial high-light-inducible proteins--protectors of chlorophyll-protein synthesis and assembly. *Biochim Biophys Acta* **1857**: 288–295.

Laslett, D., and Canback, B. (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**: 11–16.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86–89.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci U S A* **101**: 11013–11018.

Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., Kettler G., Sullivan M.B., Steen R., Hess W.R., Church G.M., Chisholm S.W. (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**: 83–86.

Liu, X., Shi, M., Kong, S., Gao, Y., and An, C. (2007) Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium *Phormidium foveolarum*: complete genome sequence and DNA translocation. *Virology* **366**: 28–39.

Liu, X., Kong, S., Shi, M., Fu, L., Gao, Y., and An, C. (2008) Genomic analysis of freshwater cyanophage Pf-WMP3 infecting cyanobacterium *Phormidium foveolarum*: the conserved elements for a phage. *Microb Ecol* **56**: 671–680.

Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A., and Alcami, A. (2009) High diversity of the viral community from an Antarctic lake. *Science* **326**: 858–861.

Mackenzie, J.J., and Haselkorn, R. (1972) Photosynthesis and the development of blue-green algal virus SM-1. *Virology* **49**: 517–521.

Mann, N.H. (2003) Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol Rev* **27**: 17–34.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Marine ecosystems: bacterial photosynthesis genes in a virus. *Nature* **424**: 741.

Martin, D., and Rybicki, E. (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16**: 562–563.

Martin, D.P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015) RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* **1**: vev003.

Middelboe, M., Jacquet, S., and Weinbauer, M. (2008) Viruses in freshwater ecosystems: an introduction to the exploration of viruses in new aquatic habitats. *Freshwater Biol* **53**: 1069–1075.

Millard, A.D., Gierga, G., Clokie, M.R., Evans, D.J., Hess, W. R., and Scanlan, D.J. (2010) An antisense RNA in a lytic cyanophage links *psbA* to a gene encoding a homing endonuclease. *ISME J* **4**: 1121–1135.

Mohiuddin, M., and Schellhorn, H.E. (2015) Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**: 960.

Muhling, M., Fuller, N.J., Millard, A., Somerfield, P.J., Marie, D., Wilson, W.H., Scanlan D.J., Post A.F., Joint I., Mann N.H. (2005) Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environ Microbiol* **7**: 499–508.

Nino-Garcia, J.P., Ruiz-Gonzalez, C., and Del Giorgio, P.A. (2016) Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *ISME J* **10**: 1755–1766.

Orellana, L.H., Rodriguez, R.L., and Konstantinidis, K.T. (2017) ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. *Nucleic Acids Res* **45**: e14.

Padidam, M., Sawyer, S., and Fauquet, C.M. (1999) Possible emergence of new geminiviruses by frequent recombination. *Virology* **265**: 218–225.

Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin E., Ivanova N.N., Kyrpides N.C. (2016) Uncovering Earth's virome. *Nature* **536**: 425–430.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Posada, D., and Crandall, K.A. (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A* **98**: 13757–13762.

Puxty, R.J., Millard, A.D., Evans, D.J., and Scanlan, D.J. (2015) Shedding new light on viral photosynthesis. *Photosynth Res* **126**: 71–97.

Rho, M., Tang, H., and Ye, Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**: e191.

Rho, M., Wu, Y.W., Tang, H., Doak, T.G., and Ye, Y. (2012) Diverse CRISPRs evolving in human microbiomes. *PLoS Genet* **8**: e1002441.

Rodriguez-R, L.M.K., and Konstantinos, T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr* **4**: e1900v1.

Rodriguez, R.L., Gunturu, S., Tiedje, J.M., Cole, J.R., and Konstantinidis, K.T. (2018) Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* **3**: e00039–00018.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahe, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**: e2584.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**: e985.

Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013) Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol* **3**: 130160.

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

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

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A. et al. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689–693.

Safferman, R.S., and Morris, M.E. (1963) Algal virus: isolation. *Science* **140**: 679–680.

Sandaa, R.A., Clokie, M., and Mann, N.H. (2008) Photosynthetic genes in viral populations with a large genomic size range from Norwegian coastal waters. *FEMS Microbiol Ecol* **63**: 2–11.

Sano, E., Carlson, S., Wegley, L., and Rohwer, F. (2004) Movement of viruses between biomes. *Appl Environ Microbiol* **70**: 5842–5846.

Sarma, T.A. (2012) Cyanophages. In *Handbook of Cyanobacteria*. Boca Raton, FL, USA: CRC Press, pp. 417–486.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D.B., Yooseph S., Zeidner G., Golden S.S., Mackey S.R., Adir N., Weingart U., Horn D., Venter J.C., Mandel-Gutfreund Y., Béjà O. (2007) Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J* **1**: 492–501.

Sherman, L.A. (1976) Infection of *Synechococcus cedrorum* by the cyanophage AS-1M. *Virology* **71**: 199–206.

Sime-Ngando, T., and Colombet, J. (2009) [Virus and prophages in aquatic ecosystems]. *Can J Microbiol* **55**: 95–109.

Smith, J.M. (1992) Analyzing the mosaic structure of genes. *J Mol Evol* **34**: 126–129.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* **4**: e234.

Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R., DeFrancesco A.S., Kern S.E., Thompson L.R., Young S., Yandava C., Fu R., Krastins B., Chase M., Sarracino D., Osburne M.S., Henn M.R., Chisholm S.W. (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ Microbiol* **12**: 3035–3056.

Suttle, C.A. (2002) Cyanophages and their role in the ecology of cyanobacteria. In *The Ecology of Cyanobacteria: Their Diversity in Time and Space*. Whitton, B.A., and Potts, M. (eds). Dordrecht: Springer Netherlands, pp. 563–589.

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A. U., Stubbe, J., and Chisholm, S.W. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proc Natl Acad Sci U S A* **108**: E757–764.

Tsementzi, D., Rodriguez-R, L.M., Ruiz-Perez, C.A., Meziti, A., Hatt, J.K., and Konstantinidis, K. (2019) Ecogenomic characterization of widespread, closely-related SAR11 clades of the freshwater genus

"Candidatus Fonsibacter" and proposal of Ca. Fonsibacter lacus sp. nov. *Syst Appl Microbiol* **42**: 495–505.

Tsementzi, D., Wu, J., Deutsch, S., Nath, S., Rodriguez-R, L. M., Burns, A.S., Ranjan P., Sarode N., Malmstrom R.R., Padilla C.C., Stone B.K., Bristow L.A., Larsen M., Glass J. B., Thamdrup B., Woyke T., Konstantinidis K.T., Stewart F. J. (2016) SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature* **536**: 179–183.

Wang, G., Murase, J., Asakawa, S., and Kimura, M. (2009) Novel cyanophage photosynthetic gene *psbA* in the floodwater of a Japanese rice field. *FEMS Microbiol Ecol* **70**: 79–86.

Wang, X., Jing, R., Liu, J., Yu, Z., Jin, J., Liu, X., Wang X., Wang G. (2016) Narrow distribution of cyanophage *psbA* genes observed in two paddy waters of Northeast China by an incubation experiment. *Virol Sin* **31**: 188–191.

Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiol Rev* **28**: 127–181.

Wilhelm, S.W., and Matteson, A.R. (2008) Freshwater and marine virioplankton: a brief overview of commonalities and differences. *Freshwater Biology* **53**: 1076–1089.

Wilhelm, S.W., Carberry, M.J., Eldridge, M.L., Poorvin, L., Saxton, M.A., and Doblin, M.A. (2006) Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of *g20* genes. *Appl Environ Microbiol* **72**: 4957–4963.

Williamson, S.J., Allen, L.Z., Lorenzi, H.A., Fadrosh, D.W., Brami, D., Thiagarajan, M., McCrow J.P., Tovchigrechko A., Yooseph S., Venter J.C. (2012) Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS One* **7**: e42047.

Wortham, B.W., Patel, C.N., and Oliveira, M.A. (2007) Polyamines in bacteria: pleiotropic effects yet specific mechanisms. *Adv Exp Med Biol* **603**: 106–115.

Xia, H., Li, T., Deng, F., and Hu, Z. (2013) Freshwater cyanophages. *Virol Sin* **28**: 253–259.

Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., and Béjà, O. (2005) Potential photosynthesis gene recombination between *Prochlorococcus* and *Synechococcus* via viral intermediates. *Environ Microbiol* **7**: 1505–1513.

Zheng, Q., Jiao, N., Zhang, R., Chen, F., and Suttle, C.A. (2013) Prevalence of *psbA*-containing cyanobacterial podoviruses in the ocean. *Sci Rep* **3**: 3207.

Zhong, X., and Jacquet, S. (2013) Prevalence of viral photosynthetic and capsid protein genes from cyanophages in two large and deep perialpine lakes. *Appl Environ Microbiol* **79**: 7169–7178.

Zhong, X., Ram, A.S., Colombet, J., and Jacquet, S. (2014) Variations in abundance, genome size, morphology, and functional role of the virioplankton in lakes Annecy and Bourget over a 1-year period. *Microb Ecol* **67**: 66–82.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Bacterial contamination of virome samples.
Fig. S2. Protein-sharing network using reference phage genomes and genome fragments recovered in this study.
Fig. S3. Relative abundance of phage VC_14LS_11 and its predicted host WB8_1B_136.
Fig. S4. Photosynthesis AMG diversity as a function of fluvial distance along the Chattahoochee River.
Fig. S5. Complete phylogenetic tree of *psbA* gene sequences
Fig. S6. Major capsid protein phylogeny of *Podoviridae* genome fragments.
Fig. S6. Phylogenetic reconstruction of photosystem-associated AMG-encoded proteins.
Fig. S7. Phylogenetic reconstruction of additional electron-transfer associated AMG-encoded proteins.
Fig. S9. Phylogenetic reconstruction light-harvesting associated AMG-encoded proteins.
Fig. S10. Alpha diversity relationship between viral and bacterial metagenome datasets.
Table S1. Viral Genome classification and AMG presence.
Table S2. vConTACT classification metadata.
Table S3. Putative host information.
Table S4. AMG diversity
Table S5. Sample metadata.