Interaction Dynamics and Virus-Host Range for Estuarine Actinophages captured by epicPCR

Eric G. Sakowski[1*], Keith Arora-Williams[1], Funing Tian[2], Ahmed A Zayed[2], Olivier Zablocki[2], Matthew B. Sullivan[2,3], Sarah P. Preheim[1*]

[1] Department of Environmental Health and Engineering, Johns Hopkins University, MD, USA

2 Department of Microbiology, The Ohio State University, OH, USA

3 Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, OH USA

* Corresponding authors

Viruses impact microbial diversity, gene flow, and function through virus-host interactions. Though metagenomics surveys are rapidly cataloging viral diversity, methods are needed to capture specific virus-host interactions *in situ*. We leveraged metagenomics and repurposed emulsion paired isolation-concatenation PCR (epicPCR) to investigate viral diversity and virus-host interactions *in situ* over time in an estuarine environment. The method fuses a phage marker, the ribonucleotide reductase (RNR) gene, with the host 16S rRNA gene of infected bacterial cells within emulsion droplets providing single-cell resolution for dozens of samples. EpicPCR captured *in situ* virus-host interactions for viral clades with no closely related database representatives. Abundant freshwater Actinobacteria lineages, in particular *Rhodoluna*, were the most common hosts for these poorly characterized viruses, with interactions correlated with environmental factors. Multiple methods used to identify virus-host interactions, including

epicPCR, identified different and largely non-overlapping interactions within the vast virus-host interaction space. Tracking virus-host interaction dynamics also revealed that multi-host viruses had significantly longer periods with observed virus-host interactions, while single-host viruses were observed interacting with hosts at lower minimum abundances, suggesting more efficient interactions. Capturing *in situ* interactions with epicPCR revealed environmental and ecological factors shaping virus-host interactions, highlighting epicPCR as a valuable technique in viral ecology.

Viruses impact the diversity and function of microbial communities from the open ocean[1] and soils[2] to the human gut[3] and are major contributors to daily bacterial mortality[4], carbon export[5] , and biogeochemical cycling[6]. In the Chesapeake Bay, the largest estuary in the United States, up to 20% of the bacterial community can be lysed by viruses per hour,[7] although environmental factors (e.g. tidal mixing), could modulate viral production and viral-mediated mortality, as observed in other ecosystems[8]. However, cultivation[9,10] and theoretical models[11,12] suggest viral pressure is unequally distributed across the microbial community, implying a subset of virus-host infections contribute disproportionately to microbial mortality, influencing microbial community diversity and biogeochemical cycling. Since individual virus-host interactions (i.e. any physical association including attachment, injection of genetic material, latent period, lysogeny) could be influenced by different ecological and environmental factors, establishing connections between virus-host interactions and environmental factors will be the first step in fine-tuning ecosystem models to better predict variability in viral productivity and bacterial mortality.

Viral metagenomics has greatly expanded our knowledge of environmental viral diversity; yet, linking viruses to their hosts remains a bottleneck for investigating the relative ecosystem impact of individual virus-host pairs. Bioinformatics approaches applied to bacterial and viral shotgun metagenomes show promise for identifying hosts for viral communities[13]. Techniques such as CRISPR spacer matches, sequence homology, and k-mer frequencies have been relatively successful, linking 35% of soil viral populations to putative hosts[2]. Floods of oceanic virus data[14] and metagenome-assembled genomes (MAGs)[15] will likely improve *in silico* host predictions for marine viruses. Still, only an estimated 10% of uncultivated bacterial lineages contain CRISPR-Cas systems[16], limiting efforts to link viruses and hosts by CRISPR spacer homology. Comparisons of mycobacteriophage suggest some viruses switch hosts too quickly for their genomes to evolve similar DNA signatures to their host[17], thus constraining *in silico* predictions via k-mer frequencies. Other techniques, like spot, plaque and liquid assays[18] and viral tagging[19,20] require cultivation of the host, preventing investigations of viral interactions for the majority of bacterial populations that cannot be readily cultivated. Cultivation-independent approaches like phageFISH[21] , polonies, or microfluidic digital polymerase chain reaction[22] are important techniques in viral ecology, but require laborious probe design and optimization, specialized expertise and equipment, or yield few positive interactions. Other culture-independent techniques like proximity ligation[23,24] and single-cell genomics[25] are exciting non-specific approaches for detecting virus-host interactions, but non-targeted approaches make consistently observing the same virus-host interactions over time inefficient. A targeted and cost-effective approach capturing virus-host interactions *in situ* using standard molecular biological equipment could fill a gap in current techniques that may reveal ecological and environmental factors influencing specific virus-host interactions.

We sought to develop such a technique by repurposing emulsion paired isolation-concatenation PCR (epicPCR)[26] to evaluate *in situ* virus-host interaction dynamics of ecologically important bacterial populations. This analysis revealed abundant freshwater Actinobacteria populations as the primary host of poorly characterized viruses sharing genetic similarity with cyanosipho- and podoviruses. Additionally, environmental factors partially explained the observed interactions between one abundant Actinobacteria population (*Rhodoluna*) and viral clades. We also identified ecological differences in "interaction life-span" and minimum host abundance between specialist and generalist viruses. These results demonstrate that this approach complements existing techniques in viral ecology.

Results

Chesapeake Bay viral populations are endemic, and seasonally dynamic

We used shotgun metagenomics to investigate temporal changes in viral populations and virus-host interactions across two years in surface water samples in the Rhode River, a tidal estuary of the Chesapeake Bay (Extended Data Fig. 1). In total, 9,392 viral populations (approximately species level genotypes) were assembled from bacterial and viral shotgun metagenomic libraries. Viral diversity in the Rhode River inlet (hereafter referred to as the Chesapeake Bay) was compared to reference database representatives (RefSeq), Global Ocean viromes [the Global Ocean Virome 2.0 (GOV 2.0) dataset[14]] and U.S. East coast estuaries (Damariscotta Salt Bay and Tampa Bay). Overall, Chesapeake Bay viral populations were distinct from reference database representatives. Chesapeake Bay viral populations formed 473 clusters in a network analysis of shared genes[27], but only 40 (8%) contained RefSeq representatives. Over 2,000 Chesapeake Bay viral populations were not assigned a single cluster.

Additionally, only 5% and 1% of viral populations could be detected in other U.S. East coast estuaries or GOV 2.0 datasets, respectively (Fig. 1). This highlights the potentially endemic nature of Chesapeake Bay viral populations and is consistent with prior inferences of endemism from cyanophage gene markers[28]. Endemism is likely attributable to the Bay's long residence time (6-7 months) promoting distinctly estuarine microbial populations[29].

Consistent with previous reports[30,31], bacterial and viral communities displayed distinct seasonal trends (Fig. 2). Winter viral communities from 2018 shared nearly three times more viral populations with a winter 2012 sample[32] than with spring and summer samples from the same year (Fig. 2b, Extended Data Fig. 2a). Likewise, spring 2017 and 2018 samples shared half of their viral populations with each other but fewer than 30% with other seasons in 2018 (Fig. 2b, Extended Data Fig. 2a), illustrating that Chesapeake Bay viral communities had significant seasonal variability ($p = 0.0003$; Extended Data Fig. 2b) but no significant inter-annual variability (Extended Data Fig. 2c).

Actinobacteria are commonly identified hosts for observed viral populations

We used bioinformatics and experimental approaches to predict hosts for Chesapeake Bay viral populations. We inferred hosts through similarity to reference sequences with the viral marker gene ribonucleotide reductase (RNR), which is broadly distributed across aquatic dsDNA lytic viruses.[33,34] In total, 634 Chesapeake Bay viral populations contained RNR genes. RNR-containing populations accounted for $5.9 \pm 0.8\%$ of viral populations and $8.9 \pm 3.2\%$ of the viral community in any given time point (Extended Data Fig. 3a,b) and captured the seasonal dynamics of the viral community ($r_s = 0.93$, $p = 0$; Extended Data Fig. 4), marking RNRs as a good proxy of aquatic viral diversity in our study. Spring RNR genes were similar to

Cyanomyoviruses, Gammaproteobacteria siphoviruses, and Alphaproteobacteria podoviruses (Extended Data Fig. 5a). The relative abundance of these myovirus and siphovirus populations decreased in winter samples, while the relative abundance of podoviruses infecting Alphaproteobacteria increased (Extended Data Fig. 5a). The summer sample was largely composed of populations with no viral reference database representatives, and overall 54 ± 19% of the RNR sequences were not classified using this homology-based approach. However, even 'classifiable' RNR sequences were only distantly related to viral reference sequences (Extended Data Fig. 5b), and homology to viral references does not necessarily indicate a common host. For example, Pelagibacter phage HTVC008M and T4-like cyanophages share 71% RNR nucleotide identity and would be classified together by our cutoffs; yet, they infect hosts from different phyla. Therefore, gene homology may lack the resolution to predict hosts for environmental viruses that are only distantly related to reference database representatives.

Cyanopodoviruses have been identified as dominant cyanophage populations in the Chesapeake Bay[30]; yet, cyanosipho-and-podoviral RNR genes from the Chesapeake Bay viral metagenomes were only distantly related to reference cyanophage sequences (Extended Data Fig. 5b). To investigate hosts associated with these RNR gene sequences, we modified an existing single-cell gene fusion protocol (epicPCR)[26] and applied it to our samples (Table S1). EpicPCR takes advantage of the close physical proximity of phage and host DNA within infected bacterial cells to fuse viral marker and host 16S ribosomal RNA (rRNA) genes within emulsion droplets, providing single-cell level resolution (Fig. 3) even without *a priori* knowledge of putative hosts. Auxiliary metabolic genes are an unusual choice for a viral marker given the possibility of horizontal gene transfer[34], but reference cyanosipho-and-podoviral RNR genes form a monophyletic clade (the Cyano SP clade[35]) distinct from other known viral and microbial

RNR genes, making Cyano SP RNR an ideal marker gene choice for epicPCR (Extended Data Fig. 3c). Our primers specifically amplified RNR genes related to those of cyanosiphoviruses and cyanopodoviruses (Table S2, Extended Data Fig. 3c). Furthermore, these amplicons were dissimilar to microbial RNR genes from our bacterial shotgun metagenomes and UniRef[36] representative sequences (Extended Data Fig. 3c), suggesting the primers specifically amplify viral RNR genes.

After confirming primer specificity, we validated epicPCR specificity using uninfected mock communities and spike-in controls. With a complex mock community spiked into an environmental sample, uninfected control cells comprised as much as 16% of the total community and four of the ten most abundant community members. Despite their dominance, mock community control sequences were not observed in any fusion products (Fig. 3b). Additionally, we ensured specificity by adding uninfected *E. coli* cells to replicate environmental epicPCR reactions, to control for non-specific interactions. Although uninfected control sequences were occasionally observed in the fusion data, these non-specific interactions were not consistently associated with any specific phage sequence. Thus, non-specific associations were removed by requiring that virus-host sequence pairs be observed in at least three libraries to be considered a positive interaction (Fig 3c).

EpicPCR yielded 8,319 fusion amplicon sequences, containing 27 unique hosts (identical 16S rRNA gene sequence), 40 unique phages (identical RNR), and 95 unique phage-host interactions (Fig. 4, Table S3). RNR genes from epicPCR formed three distinct phylogenetic clades (Fig. 4a). These genes shared greatest homology with T7-like Chesapeake Bay cyanopodoviral isolates S-CBP1, S-CBP3, and S-CBP4 but had less than 80% nucleotide identity to any reference sequence (Table S2). Recruiting viral metagenome reads to the RNR amplicons

demonstrated these viral RNR sequences were not abundant in our samples, which highlights the sensitivity of this method (Extended Data Fig. 6).

Unexpectedly, these RNR sequences were most frequently associated with Actinobacteria (Fig. 4) and are hereafter referred to as CSP-like Actinophage. Approximately 80% (31/40) of the RNR sequences were linked to a single host (Actinobacteria member *Rhodoluna*), identifying *Rhodoluna* as the primary host for these cyanopodoviral-like populations. We looked for other cyanophage genes (e.g. *psbA*) on the same contig as these genes to investigate their genomic context. However, RNR sequences failed to assemble into longer contigs, likely due to their low abundances and micro-diversity[37]. Other observations provide indirect evidence that these CSP-like Actinophage populations differ from their cyanophage counterparts. From May to December 2018, Chesapeake Bay *Rhodoluna* and CSP-like Actinophage abundances (Extended Data Fig. 3d) varied more similarly to each other (32 and 31-fold, respectively) than to cyanopodoviruses (8-fold change) and Cyanobacteria (15-fold change). This suggests CSP-like Actinophage abundances are associated with *Rhodoluna*, rather than with Cyanobacteria. After screening single-amplified genome (SAG) libraries, no CSP-like Actinophage RNR gene sequences could be amplified from ~300 Cyanobacteria SAGs. In contrast, one CSP-like Actinophage RNR gene sequence was associated with an Alphaproteobacteria SAG, while a second was associated with an Acidimicrobiia (phylum Actinobacteria) SAG library (Fig. 4a), providing further support that poorly characterized CSP-like Actinophage interact with Actinobacteria.

Actinobacteria were also commonly predicted hosts of Chesapeake Bay viral populations using *in silico* approaches. We applied a variety of bioinformatics approaches to identify hosts for observed viral populations. Markov model-based prediction resulted in significant ($p < 0.05$) host predictions for about half of the viral populations from either reference genomes or

metagenome assembled genomes (MAGs; Extended Data Fig. 7). Four of the ten most frequently predicted hosts were Chesapeake Bay MAGs, all of which were classified as Actinobacteria. Among Chesapeake Bay MAGs, Actinobacteria were the predicted hosts for more viral populations (57%) than all other MAGs combined even though Actinobacteria only comprised 25% of MAG classifications overall (Extended Data Fig. 7, 8). 2,204 tRNA sequences were found in the viral populations, but this resulted in only 40 perfect matches between viral populations and MAGs or reference genomes to allow for host inference. A majority of MAGs harbored at least one CRISPR spacer, although only three MAGs had CRISPR spacers matching observed viral populations. A comparison between bioinformatics methods demonstrates that each approach identifies different and largely non-overlapping interactions within the vast virus-host interaction space (Extended Data Fig. 9). However, Actinobacteria were consistently observed as putative hosts of viral populations in the Chesapeake Bay with all methods except CRISPR analysis, suggesting significant viral pressure on Actinobacteria populations in this environment throughout the year.

We also investigated Chesapeake Bay microbial susceptibility to viruses *in vitro* by comparing growth of bacterial populations in incubations of water samples with and without active viruses (Extended Data Fig. 10). Alphaproteobacteria were the most frequently identified susceptible bacterial populations, and Flavobacteriia displayed an interesting split between the susceptible Flavobacteriaceae and the resistant Cryomorphaceae (Extended Data Fig. 10c). Actinobacteria lineages Acidimicrobiia and Actinobacteria (class level) were also among the most commonly identified susceptible taxa (Extended Data Fig. 10c). Three *Rhodoluna* spp. were among these susceptible populations, including the primary *Rhodoluna* host identified by epicPCR and a population that differed from it by one nucleotide (Extended Data Fig. 10c).

Previously, Actinobacteria was the most frequently identified host in an analysis of 2,000 freshwater phage genomes.[38] Although these phages lacked homology to the Chesapeake Bay amplicons, this suggests that Actinobacteria may be under substantial viral pressure across both freshwater and estuarine environments. More work is needed to understand how viruses shape Actinobacteria abundance, diversity, evolution and metabolic processes.

In situ *viral-host interaction dynamics reveal ecologically differentiated viral clades*

EpicPCR results suggest RNR viral clades are ecologically differentiated. Interactions between Clade I CSP-like Actinophages and their hosts peaked in May before largely disappearing during the summer (Fig. 4B). Clade I phage sequences were most frequently associated with multiple hosts (4/6 sequences) and the greatest number of hosts (Fig. 4A, C). Host ranges for Clade I phages were significantly correlated with CSP-like Actinophage abundances throughout 2018 ($r_s = 0.69$; $p = 0.002$). Thus, broader host ranges of Clade I phages could be a byproduct of increased contact rates resulting from higher viral abundances. Broader host ranges may also provide an advantage for late spring viral populations under dynamic conditions characterized by large freshwater influxes.

In contrast, Clade II and III phage-host interactions were primarily observed during the summer (Fig. 4B). Most (7/8) Clade II phage sequences were associated with a single host, while clade III sequences were split between single (11/19) and multiple (8/19) hosts. Neither clade displayed broad host ranges or correlations between host range and viral abundance like the Clade I phage sequences. Interactions between *Rhodoluna* and Clade II and Clade III phage populations alternated at most time points (Fig. 5), which may be a reflection of tidal influences on phage and/or host diversity. Supporting this hypothesis, Clade III interactions were negatively

correlated with fluorescent dissolved organic matter (r_s = 0.7; p = 0.003), influenced by tide at our sample site[39], and positively correlated with salinity (r_s = 0.6; p = 0.02) (Table S4). Although mean tidal amplitude in the Rhode River is 0.3 m, water level is strongly influenced by weather conditions[40], which might explain the lack of correlation with water level. This suggests Clade II and III phage populations may represent riverine and estuarine diversity, respectively. However, we cannot rule out other possible explanations, such as changes to host physiology with different environmental conditions[41] or antagonistic evolution between phage and host[42].

Host range associated with virus-host interaction lifespan and efficiency

Interactions captured by epicPCR illuminated differences in *in situ* interaction persistence and interaction efficiency between single-host (specialist) and multi-hosts (generalist) phage. Generalist phages were observed across a significantly greater number of sample dates (4.5 ± 1.9 weeks) than specialist phages (2.5 ± 1.0 weeks; p = 0.002; Fig. 6A). This suggests that broader host range increases the 'interaction lifespan' of phage, a possible evolutionary advantage in highly dynamic systems where viral infections appear to be largely ephemeral.[43,44]

EpicPCR and 16S rRNA marker gene analyses also revealed differences in the minimum abundance of a host when it was associated with generalist or specialist phages. Virus-host interactions could be detected by epicPCR at significantly lower host abundances for specialist than generalist phages (p = 0.001; Fig. 6B). The minimum host abundance across time-points where the host was associated with generalist CSP-like Actinophage ranged from 0 – 2.1% of the community (0.63% mean minimum abundance). By comparison, the minimum abundance of hosts associated with specialist CSP-like Actinophage ranged from 0 – 0.33% of the community (0.11% mean minimum abundance). Thus, epicPCR was capable of detecting virus-host

interactions even when the host went undetected in the 16S rRNA gene libraries. Five of the eight hosts associated with specialist phages were also associated with generalist phages, suggesting that host identity was not inherent to these observed threshold differences. Instead, observed minimum host abundance differences for specialist and generalist phages could be attributed to increased infection efficiencies among specialist phages compared to generalist phages. Supporting these findings, previous work identified fitness costs associated with increased host range in cultivated phage-host systems[45], and trade-offs in viral infection efficiency with increasing host range[11]. Considering these differences, there appears to be a trade-off between 'interaction lifespan' and infection efficiency for CSP-like Actinophages in the Chesapeake Bay.

Discussion

EpicPCR can complement existing methods in viral ecology by linking viral and bacterial populations *in situ*. EpicPCR targets specific viral populations, requires little more than standard molecular biological equipment, and is cost-effective, making it ideally suited to probe specific virus-host interactions with high temporal and genetic resolution. This targeted approach is more cost-effective than non-specific approaches like proximity ligation or single-cell genomics when studying a specific viral target. EpicPCR can complement large-scale metagenomic sequencing efforts[6,46] through host identification for uncultivated viruses, a notable advantage over cultivation-based approaches, like viral tagging[47]. This method does not require cultivation to reveal virus-host associations. However, as the dominant host was closely related to cultivated Actinobacteria species (i.e. *Rhodoluna*), future work is needed to demonstrate the potential of this technique to reveal virus-host relationships in uncultivated lineages. Digital PCR and single-

276 cell genomics can currently only accommodate hundreds to thousands of reactions and require
277 costly and/or specialized equipment. In contrast, epicPCR efficiently screens hundreds of
278 thousands of reactions within an emulsion, can be performed using standard molecular biology
279 equipment, and efficiently identifies positive associations when the vast majority of reactions are
280 expected to be negative. Although primer bias remains an issue for epicPCR, any of the
281 commonly applied viral marker genes could be used[48]. Primers could be developed to target
282 abundant viruses, such as vSAG 37-F6[49,50], or verify host predictions, such as potential archaeal
283 viruses[51]. Observations from epicPCR could test underlying assumptions of viral-host infection
284 networks predicted from metagenomic time series.[52] Unlike cultivation-based methods or *in*
285 *silico* predictions, epicPCR can capture virus-host interactions when and where they occur,
286 revealing relationships with ecological and environmental factors. While epicPCR captures close
287 physical contact, it does not confirm active infections or differentiate between lytic or lysogenic
288 infections, similar to limitations of single-cell genomics[25]. Pairing epicPCR with viral marker
289 gene expression[43] could confirm active infections. Host range capabilities of epicPCR paired
290 with quantitative methods of viral abundance, such as qPCR[53], ddPCR[50] or polonies[54], could be
291 used to estimate the impact of viruses on different microbial taxa.

292 In this study, we developed primers targeting a well-studied group of cyanophages, which
293 yielded unexpected host associations and interaction dynamics. Our analysis revealed that these
294 CSP-like RNR populations commonly interact with Actinobacteria, which was not predicted
295 from homology searches of phage associated with cultured Actinobacteria, such as within
296 PhagesDB[55]. Associations between observed viral populations and Actinobacteria were robust
297 across methodologies, with all but one technique predicting Actinobacteria as a common host for
298 observed viral populations. In addition, some phages interacting with Actinobacteria were also

299 associated with hosts from other phyla. Although most viruses are believed to have narrow host
300 ranges, host enrichment may select against broad-host range phages[56], and cultivated viruses
301 capable of infecting hosts spanning taxa at the order level have been reported[57]. Additionally,
302 Paez-Espino et al. (2016)[58] found 1% of viruses from metagenomic libraries had predicted hosts
303 spanning phyla based on CRISPR spacer and tRNA homology, including phages associated with
304 Actinobacteria. Furthermore, cultivation-based host range experiments often screen against
305 related hosts. We most frequently observed generalist phages associated with *Rhodoluna*
306 (Actinobacteria) and Halieaceae (Gammaproteobacteria), which would be unlikely to be
307 screened together. These broad host ranges could represent artifacts of the epicPCR method even
308 though we used controls in all environmental samples to remove spurious interactions. However,
309 observations of multiple different phages associated with the same hosts (*Rhodoluna* and
310 Halieaceae) suggest this is unlikely.
311
312 ***Conclusions***
313 Viruses influence biogeochemistry through interactions with hosts mediating key
314 microbial processes. Using a combination of shotgun metagenomics, marker gene analysis,
315 dilution experiments and a unique gene fusion technique, we investigated factors that influence
316 virus-host interactions within the largest estuary in the United States. We found Actinobacteria
317 populations are under substantial viral pressure within this ecosystem and their interactions with
318 specific phage clades seem to vary with tidally-influenced environmental factors. Virus-host
319 interactions revealed by epicPCR suggest that single- and multi-host phages have different
320 'interaction lifespans' and host abundance characteristics. If similar results are found broadly
321 across diverse viral lineages, these characteristics could represent an important trade-off shaping

viral evolution and might warrant separate parameters for generalist and specialist viruses in ecological models. This combination of bioinformatics and experimental approaches provided high genetic and temporal resolution of viral interactions with one of the most abundant heterotrophic bacterial populations within this ecosystem that can be widely applied across aquatic ecosystems to gain insight into viral ecology.

Materials and Methods

Sample Collection and preservation

Surface water samples were collected from May 2017 and May to December 2018 from the mouth of the Rhode River, a tidal estuary on the Western Shore of the Chesapeake Bay (Edgewater, MD), off of the Smithsonian Environmental Research Center research pier (38.89 N 76.54 W). Samples were collected five times a day over three days from 05/16/17-05/18/17, between 10:30 am and 4:30 pm. Samples were also collected at 12:00 pm weekly from 5/24/18 to 8/9/18, on 8/23/18, then again weekly from 12/6/18 to 12/27/18. Briefly, 25 mL of water sample was combined with 25 mL of 50% (v/v) sterile glycerol. Glycerol samples were stored on dry ice for transport back to Baltimore, MD and subsequently stored at -80°C until processing. Approximately 120 mL of water sample was also collected per time point and filtered through a 0.2 μm PES membrane filter (Millipore, Inc.) for bacterial shotgun metagenomic analyses. Filters were stored on ice for transport back to Baltimore, MD and kept at -80°C until processing. Viral filtrates (< 0.2 μm fraction) were collected and incubated with $FeCl_3$ as previously described [59] during transport back to Baltimore, MD. Viral filtrates were filtered through a 0.2 μm PES membrane filter (Millipore, Inc.) post-incubation, and filters were stored in the dark at 4°C until processing for shotgun sequencing. Water conditions were recorded from

345 the continuous water monitoring station located at Smithsonian Environmental Research Center
346 (http://nmnhmp.riocean.com, Table S5).
347
348 *Bacterial and viral gene counts*
349 Bacterial and viral abundances were estimated by quantitative PCR. To create a standard
350 curve of gene copy number, viral 'Cyano SP'-like RNR genes were amplified from
351 environmental samples using primers Cyano_II_F and Cyano_II_R (0.3 uM final concentration
352 each, Table S6). Amplicons were run on a 1.5% agarose gel and visualized with SYBR Safe
353 DNA gel stain (Invitrogen, 1x final concentration). RNR bands were cut out, gel purified (Zymo,
354 Inc.), and cloned into chemically competent *Escherichia coli* cells using the Zero Blunt PCR
355 Cloning Kit (Thermo Scientific) following the manufacturer's protocol. Cells were grown
356 overnight on LB + kanamycin (50 μg mL^{-1}) plates at 37°C and colonies were picked for
357 subsequent testing. Picked colonies were tested for amplification of the RNR gene. One colony
358 was serially diluted and grown on plates to determine gene copy number per dilution. The serial
359 dilution series was used to create a standard curve for use in qPCR. All standards and
360 environmental samples were run in triplicate. Three microliters of sample were combined with
361 UltraPure molecular grade water (Thermo, Inc.), SsoAdvanced Universal SYBR Green
362 Supermix (1x final concentration, Bio-Rad Laboratories, Inc.), Cyano_II_F primer (0.3 μM final
363 concentration), and Cyano_II_R primer (0.3 μM final concentration) to a final volume of 25 μL
364 (see Table S6 for primer sequences and citations). Samples were amplified on a CFX96 Real-
365 Time PCR Detection System (Bio-Rad Laboratories, Inc.) with the following conditions:
366 denaturing at 98°C for 10 minutes; 45 cycles of denaturing at 98°C for 10 seconds, annealing at
367 52°C for 30 seconds, and extension at 72°C for 45 seconds; and a final extension of 72°C for 5

minutes. The same dilution series was used to create a standard curve for the 16S rRNA gene. 16S rRNA gene counts were assessed as above with primers PE_16S_U515F and 16S_1114R (Table S6).

Shotgun metagenomic library preparation, sequencing, and processing

To investigate bacterial diversity, we employed short-read shotgun metagenomics to the bacterial fraction (> 0.2 µm). DNA was extracted from 0.2 µm filters for shotgun sequencing from water samples collected on the following dates: 05/17/17, 05/18/17, 05/19/17, 05/31/2018, 06/28/18, 08/02/18, and 12/6/18. Additionally, two positive controls were processed, a Zymo positive control community (Zymo Research) and *E. coli,* and one negative control (water). DNA extraction was performed with the DNeasy PowerWater kit (Qiagen) following the manufacturer's protocol with the following amendment: 20 µL of proteinase K was combined with 1 mL of solution PW1 in the bead tube. The bead tube was incubated at 65°C for ten minutes prior to bead beating. Libraries were prepared with the Nextera DNA Flex Library Prep kit (Illumina, Inc.) following the manufacture's protocol and sequenced on an Illumina MiSeq (2 x 300 bp) at the Genetic Core Research Facility at Johns Hopkins University.

Sequences were quality filtered and trimmed with trimmomatic (v. 0.38)[60], assembled with metaSPAdes (v. 3.13.1)[61]. The assembly was used to generate metagenome assembled genomes (MAGs) through metaWRAP (v1.2)[62] with metabat[63] and maxbin2[64]. Binning_refiner[65] was used to create the final set of MAGs with at least 50% completeness and less than 10% contamination, as determined by CheckM[66] were also run in metaWRAP. MAGs were classified using GTDB-Tk (v1.1.0) [67].

16S rRNA gene amplicon library preparation, sequencing, and processing

16S rRNA genes were amplified from 2017 and 2018 Chesapeake Bay surface water samples (see Table S1) in a 25 μL PCR reaction with the following conditions: three microliters of column-purified DNA were combined with UltraPure molecular grade water (Thermo, Inc.), 10X buffer (1x final concentration), dNTPs (0.1mM each final concentration), 16S forward primer 27F (0.3 μM final concentration), 16S reverse primer PE_16S_V4_E786_R (0.3 μM final concentration), bovine serum albumin (0.02 mg/mL final concentration), and Phusion High-Fidelity DNA Polymerase (0.5U; New England BioLabs, Inc.; see Table S6 for primer sequences and citations). PCR reactions were combined with 150 μL of 4% UMIL EM90 oil (4% UMIL EM90 oil, 0.05% TritonX-100 v/v in mineral oil; Universal Preserv-A-Chem, Inc.) and emulsified by vortexing at max speed (~2,700 rpm) for one minute on a Vortex Genie 2 (MoBio). Emulsions were loaded as 50 μL aliquots and amplified with the following conditions: denaturation at 94°C for 3 minutes; 33 cycles of denaturation at 94°C for 10 seconds, annealing at 54°C for 30 seconds, and extension at 72°C for 45 seconds; and a final extension of 72°C for 5 minutes (C1000, BioRad Labs., Inc.). Samples were immediately removed upon completion of amplification and stored at -20°C until the emulsion was broken.

PCR oil emulsions were broken with isobutanol as previously described[68]. Briefly, PCR aliquots were pooled in a 1.5mL microcentrifuge tube and combined with 100 uL of sterile 5M NaCl solution and 1 mL of isobutanol. Samples were vortexed briefly to mix and centrifuged at 16,000 x g for 1 minute. The bottom aqueous layer was retained, and DNA was purified by spin column purification (Zymo, Inc.). DNA was eluted in 20 uL of Tris-HCl and stored at -20°C.

Purified DNA was run on a 1.5% agarose gel (UltraPure Agarose, ThermoFisher Scientific) and visualized with SYBR Safe DNA gel stain (Invitrogen, 1x final concentration).

414 The gel was run in 1X TBE buffer (Alfa Aesar) at 4 V/cm. 16S rRNA gene bands were
415 visualized under blue light excitation, extracted, and gel purified (Zymo, Inc.) Purified DNA was
416 eluted into 20 μL of Tris-HCl and stored at -20°C until further processing.
417 Barcodes and Illumina adapters were added to 16S rRNA gene amplicon products in two
418 subsequent limited PCR steps. Barcodes were added as follows: two microliters of purified DNA
419 were combined with UltraPure molecular grade water (Thermo, Inc.), 10X buffer (1x final
420 concentration), dNTPs (0.1mM each final concentration), 16S forward primer
421 PE_16S_V4_U515F (0.3 μM final concentration), 16S rRNA gene reverse primer with 8-mer
422 barcodes PE_IV_XXX (0.3 μM final concentration), and Phusion High-Fidelity DNA
423 Polymerase (0.5U; New England BioLabs, Inc.; see Table S6 for primer sequences and
424 citations). Samples were amplified with the following conditions: denaturing at 98°C for 30
425 seconds; 8 cycles of denaturing at 98°C for 10 seconds, annealing at 54°C for 30 seconds, and
426 extension at 72°C for 45 seconds; and a final extension of 72°C for 5 minutes. DNA was purified
427 by spin column purification (Zymo, Inc.) and eluted into 20 μL Tris-HCl. Illumina adapters were
428 then added as above with the following primers: Illumina adapter forward primer PE-III-PCR-F
429 (0.3 μM final concentration) and Illumina adapter reverse primer Barcode_Rev (0.3 μM final
430 concentration) (see Table S6 for primer sequences and citations). Samples were amplified with
431 the following conditions: denaturing at 98°C for 30 seconds; 5 cycles of denaturing at 98°C for
432 10 seconds, annealing at 54°C for 30 seconds, and extension at 72°C for 45 seconds; and a final
433 extension of 72°C for 5 minutes. DNA was purified by spin column purification (Zymo, Inc.)
434 and eluted into 20 μL Tris-HCl.
435 16S rRNA gene amplicon products were quantitated on a Qubit 3.0 fluorometer
436 (Invitrogen) and three nanograms of DNA pooled per sample for sequencing. 16S rRNA gene

amplicon libraries were sequenced on an Illumina MiSeq (2 x 300 bp) at the Genetic Core Research Facility at Johns Hopkins University. Sequence reads were processed in QIIME2[69] using the DADA2 de-noising pipeline with the following parameters: trim left = 23 bases, truncate = 200 bases, minimum fold-change of parent over abundance for chimera detection = 10. Taxonomic assignment was performed in QIIME2 with the Greengenes[70] database.

Virome shotgun metagenomic library preparation, sequencing, and processing

To investigate viral diversity, we employed short-/long-read hybrid metagenomic approaches[71] to the viral fraction (< 0.2 μm) that had been incubated with $FeCl_3$ and filtered as described above (*Sample collection and preservation* section). Samples from the following dates were used for virome short-read shotgun metagenomic analysis: 05/17/17, 05/18/17, 05/31/18, 08/02/18, 12/14/18, 12/20/18. The two samples collected on 12/14/18, 12/20/18 were also long-read sequenced. After $FeCl_3$ incubation and filtration onto 0.2 μm filters, viruses were resuspended from filters with an ascorbic acid buffer as previously described[59]. Following resuspension, viral particles were purified by cesium chloride gradient centrifugation [72] and DNA extracted with Wizard Prep Columns (Promega, Corp.). Viral metagenome short-read libraries were prepared using the NexteraXT kit (Illumina, Inc.) following the manufacturer's protocol. For samples with > 0.16 ng/μL, samples were amplified with undiluted amplicon target mix (ATM) at 15 cycles; for those with 0.1-0.16 ng/μL, samples were amplified with a 1:5 dilution of ATM at 18 cycles; for <0.1ng/μL, samples were amplified with a 1:10 dilution of ATM at 20 cycles. Short reads for all virome libraries were sequenced on an Illumina NovaSeq S4 with 75M (target) 2x150bp reads at the JP Sulzberger Genome Center (Columbia University, New York, NY). Additionally, the long-read libraries from December 2018 were prepared using

460 phenol:chloroform extraction protocol (dx.doi.org/10.17504/protocols.io.6cbhasn) and libraries
461 were prepared as previously described [71] with modifications and sequenced with an Oxford
462 Nanopore MinION instrument on a FLO-MIN106D R9 version Spot-ON flowcell (Rev D) at the
463 Ohio State University according to the manufacturer's instructions.
464 Short reads were cleaned and quality-trimmed with bbduk[73]; adapters, sequencing
465 artifacts, and PhiX sequences were removed (ktrim=r; k=23 mink=11; hdist=1; hdist2=1). Reads
466 were then quality-trimmed from both ends to remove bases with low quality scores (qtrim=rl;
467 trimq=20). Reads shorter than 30 bp (minlength=30), with Ns (maxns=0), or with an average
468 quality below 20 (maq=20) were discarded. The cleaned reads from each sample were then
469 independently assembled with metaSPAdes (v. 3.13.1)[61] using k-mer sizes: 21, 33, 55, 77, and
470 –meta parameter. Long-reads were basecalled with Guppy v.2.3.1 (Manufacturer's tool) and
471 individual barcoded sample libraries were demultiplexed with the 'barcoder' function of Guppy.
472 Long-reads were quality controlled with NanoFilt [74], in which reads were filtered by quality
473 score (Q-score ≥ 10) and minimum length (≥1kb), and finally 'headcropped' by 50bp to ensure
474 no remaining barcode sequence remained. These cleaned long-reads were used in two separate
475 assembly scenarios. First, long-reads were assembled with Flye[75] in metagenomic mode (--meta).
476 Error-correction of the Flye assemblies where performed with Pilon v.1.23 [76], using the
477 corresponding short-reads mapping information [read recruitment was performed with BWA
478 v.0.7.17 [77]] to detect and reduce basecalling errors. Second, hybrid assembly, using both long-
479 and short-reads was performed with SPAdes (v.3.13.1)[61]. Long reads were assembled through
480 hybrid assembly (with metaSpades hybrid option) and Flye[75]. In order to predict viral contigs in
481 the assembled datasets, assemblies were run through VirSorter[78] (v. 1.0.5) in the -virome mode
482 after upgrading its database with an expanded profile HMM database of viral proteins, mainly

from the GOV2.0 dataset[14]. Contigs that were resolved as categories 1, 2, 4 and 5 were retained, and filtered by length ≥ 5kb (≥1.5kb, if circular). DeepVirFinder [79] was another tool for rescuing additional viral contigs. We considered high-confidence viral contigs from DeepVirFinder to be those of scores of ≥ 0.9 with a $p \leq 0.05$, and lengths ≥ 5kb. These two sets of predicted viral contigs (from both long- and short-read assemblies) resulted in 9,392 viral populations (approximately species level genotypes) dereplicated from 14,848 virome (< 0.2μm) and metagenome (> 0.2 μm) viral contigs using "ClusterGenomes" [80] using 95% average nucleotide identity over 80% coverage of the shorter contig length for all contigs > 5 kb.[14,81-83]

To calculate the coverage of these viral populations, clean reads were mapped to the Chesapeake Bay viral population database with Bowtie2 [84] in the non-deterministic and sensitive mode. The output bam files were parsed using BamM (https://github.com/Ecogenomics/BamM) to only keep the reads that covered over 70% of the viral contig length, with over 75% read alignment length. Pysam v0.8.5 (https://github.com/pysam-developers/pysam) was then used to filter out reads with <95% identity. Trimmed pileup coverage "tpmean" for each contig was calculated using BamM and then they were adjusted for each sample by metagenome size. The same read mapping strategy was employed for RNR sequences upon constructing the rank abundance curves in Extended Data Fig. 5.

To study seasonal diversity of Chesapeake Bay viral populations, all viromes were randomly subsampled to 15M reads using bbmap "reformat" [73] with default parameters. The subsampled read libraries were assembled using metaSPAdes (v. 3.13.1) as above. The resulting assemblies were then processed with VirSorter and DeepVirFinder as above. Viral contigs extracted using the same cutoffs as mentioned before were grouped into viral populations if they

shared ≥ 95% nucleotide identity across ≥ 80% of the shorter contig length. Subsequently, the subsampled reads were recruited to the viral populations' representatives with Bowtie2, and the same read-mapping cutoffs discussed above were applied to calculate sequence-depth adjusted coverage. Taxonomic assignment of viral populations was performed using vConTACT2 [80,85] with default parameters. First, the full proteome of every viral population representative was predicted using Prodigal (v2.6.3)[86]. The protein set was then combined with all the proteins from the phage and archaeal viruses in the NCBI RefSeq v88 release. The combined set of proteins was then used as input for vConTACT2 to compute protein similarity overlaps (i.e. protein clusters), and subsequently refined into genus-level equivalent 'viral clusters' (VCs). Using this method, viruses are classified at the genus level. The resulting cluster file were imported and visualized in Cytoscape 3.7.2[87].

Seasonal dynamics and diversity analyses

Seasonal bacterial communities were assessed from 16S rRNA gene amplicon libraries. Hierarchical clustering was performed in QIIME[88] with Bray-Curtis dissimilarity. Libraries were sub-sampled at the smallest library size (150,000 observations) with ten jackknifed replicates. Community structure was visualized using heatmap.2 in the gplots[89] package in R.

Seasonal Chesapeake Bay viral communities were characterized by mapping reads to assembled contigs using Bowtie2[90] with default sensitive conditions. To compare viral populations with a deeply sequenced virome library collected in 2012 at the same location[32], 2017 and 2018 Chesapeake Bay virome contigs > 5kb were combined with contigs > 5kb from the 2012 virome that were previously assembled. These merged datasets were then dereplicated as described above, resulting in 21,784 viral populations. The Chesapeake Bay 2017 and 2018

529 reads were then mapped to this larger pool of viral populations. Only reads that mapped to
530 contigs with ≥ 90% identity over ≥ 90% of the read were retained. Viral populations were
531 considered present in the sample if retained reads mapped to ≥ 75% of the contig. In total, 10,858
532 of the 21,784 viral populations from the combined 2012, 2017, and 2018 samples were identified
533 in the 2017 and 2018 samples. Trimmed pileup coverage values were used as proxies of
534 observation counts for each population. Hierarchical clustering was performed in QIIME with
535 Bray-Curtis dissimilarity. Libraries were sub-sampled at the smallest library size (300,000)
536 observations with ten jackknifed replicates. Community structure was visualized in R as
537 described above.
538 Shannon diversity was calculated in QIIME for libraries with corresponding virome
539 samples (5/17/2017, 5/31/2018, 8/2/2018, 12/14/2018, 12/20/2018). Values were recorded as the
540 average of ten jackknifed replicates at a sampling depth of 10,000 observations for both 16S and
541 viral populations.
542 The distribution of Chesapeake Bay viral populations in ocean viral metagenomes was
543 investigated by mapping viral metagenome reads from Global Ocean Virome 2.0 database
544 libraries[14] to the Chesapeake Bay > 5kb contigs as described above for the Chesapeake Bay 2012
545 viral metagenome sample. Additional libraries from previously published freshwater and
546 estuarine environments were similarly queried[91-93], in addition to a viral metagenome from the
547 Damariscotta River Estuary, ME, USA (BioProject Accession No. PRJNA357591).
548 Fold-change in specific lineages between May and December 2018 were determined with
549 read recruitment (viral populations), qPCR (RNR genes), or a combination of qPCR and 16S
550 rRNA gene libraries. The difference in cyanopodovirus abundance (8-fold change) was
551 determined through read recruitment to populations classified as cyanopodovirus. The difference

552 in Cyanobacteria (15-fold change) and *Rhodoluna* (32-fold change) abundance was determined
553 by using the relative abundances from 16S rRNA gene libraries and the absolute abundance of
554 16S rRNA genes with qPCR. The difference in SCP-like Actinophage abundance (31-fold
555 change) was determined with qPCR.
556
557 *Statistical analysis*
558 The Kruskal-Wallis test for non-normal distributions was used to determine significance
559 of viral seasonal dynamics ($p = 0.0003$, Fig. S2b) and the growth of individual bacterial
560 populations in *in situ* incubations in the presence and absence of active viruses (FDR $p < 0.05$,
561 Extended Data Fig. 10). Spearman's Rho for non-normal distributions was used to determine the
562 p-value of the linear regression between the diversity of RNR genes with overall viral diversity
563 (linear regression $r^2 = 0.95$, $p = 0$, Extended Data Fig. 4), the correlation of Clade I phage with
564 qPCR measurements of CSP-like Actinophage abundances (Spearman's Rho = 0.69; $p = 0.002$),
565 the correlation of Clade III with fluorescent dissolved organic matter (Spearman's Rho = 0.7; $p =$
566 0.003; Table S4) and salinity (Spearman's Rho = 0.6; $p = 0.02$; Table S4). The interaction
567 lifespan ($p = 0.002$; Fig. 6A) and minimum host abundance ($p = 0.001$; Fig. 6B) differences
568 between generalist and specialist phage were determined with a two-tailed T Test for normally
569 distributed data. The calculation of minimum abundance of hosts associated with generalist
570 phage ranged from 0 – 2.1% of the community (0.63% mean minimum abundance, $n = 21$),
571 while the calculation of minimum abundance of hosts associated with specialists ranged from 0 –
572 0.33% of the community (0.11% mean minimum abundance, $n = 8$). The host was undetectable
573 in 5 out of 29 cases, but omitting these observations had no effect on the analysis.
574

575 *Ribonucleotide reductase homology analysis*

576 Open reading frames (ORFs) were called for all Chesapeake Bay > 5kb contigs using
577 MetaGeneMark[94]. RNR alpha subunit genes were identified by BLASTx query of ORFs (e-value
578 < 1E-10) against a database of Uniref50[95] RNR cluster representative sequences. Putative RNR
579 genes were translated to amino acid sequences and aligned with MAFFT (v. 7.453)[96]. Aligned
580 sequences were visually inspected in Geneious[97] v. 9.1.5 for the presence of conserved catalytic
581 residues C439, E441, C462 indicative of RNR proteins (amino acid positions of *Escherichia coli*
582 nrdA gene product). Only sequences containing these conserved amino acids and spanning the
583 C462 to P621were retained. All retained sequences were queried against viral sequences in the
584 NCBI nr database by BLASTn to identify putative viral and host taxonomy. Top hits with a
585 percent identity ≥ 65% across at least 90% of the query sequence were recorded. All sequences
586 with top hits below this threshold were reported as 'Unclassified'.

587 Chesapeake Bay RNR amplicon sequences identified as CSP-like Actinophage by
588 epicPCR were queried against 2,034 freshwater phage genomes by BLASTn with an e-value
589 cutoff of 0.001. None of the freshwater phage genomes shared homology with the RNR
590 amplicon sequences.

591

592 *Bioinformatics host prediction*

593 Markov-model based predictions with WIsH[98] was used to determine potential hosts for
594 viral contigs greater than 10 KB in length. The entire dataset used in the WIsH analysis included
595 viral contigs and metagenome assembled genomes (MAGs >50% completion, <10%
596 contamination) from the Rhode River, along with the viral and hosts reference genomes from the
597 WIsH benchmark dataset, downloaded from NCBI. Because an appropriate null model is not

598 available for this unique estuarine environment, we assume that for every bacterial model, the set
599 of phage genomes in the dataset for which it is a host is negligible compared to the set of phage
600 genomes that for which it is *not* a host. Reference host and viral genome relationships in the
601 benchmark dataset were used to verify the performance under these assumptions. Running the
602 benchmark dataset with the null model yielded similar results as those previously reported (58%,
603 as compared to 63%, accuracy reported at the genus level). Thus, we applied the same
604 assumption for the null model to the analysis of environmental data. The analysis required top
605 hits to have a likelihood value in the top 5% of all calculated likelihoods ($p <= 0.05$). Of the
606 9,392 viral contigs from the Chesapeake Bay, almost half had top predictions that were within
607 the top 5% of all tested likelihoods. Of significant hits, 80% of top predictions of hosts for
608 environmental viral population mapped to reference genomes and 20% of significant top
609 predicted hosts were MAGs.

610 Hosts for viral populations were determined by transfer RNA (tRNA) homology search.
611 tRNA genes were predicted from the 2017 and 2018 viral populations using ARAGORN
612 (v1.2.38)[99]. Matching tRNA sequences in either the MAGs or GenBank (nt database downloaded
613 on Nov. 6, 2020) was determined by BLAST[100] with 100% identity and coverage. The host for
614 each viral population was taken from either a match to a viral sequence, in which case the host of
615 the matching virus was assumed as the host, or a matching bacterial sequence. NCBI taxonomy
616 was used to get genus and phylum level taxonomic information for each match.

617 Potentially interacting viral populations were also determined for MAGs through
618 CRISPR spacer matches to the observed viral populations. CRISPRs were identified within each
619 metagenome assembled genome with CRISPRFinder online[101] using default settings. BLAST[100]
620 was used to identify putatively interacting viral populations from the viral populations with

100% identity and coverage. Self-self matches were removed when contigs were identical between datasets.

epicPCR of environmental samples

Glycerol samples were thawed on ice and one mL was added to three replicate 1.5 mL microcentrifuge tubes per sample. One replicate was left unamended, while the other two were spiked with *E. coli* to approximately 0.1% and 1% of the bacterial community, respectively, to identify false-positive interactions. A fourth replicate with 5% *E. coli* was processed for seven of the time points. Samples were centrifuged at 25,000 x g for 10 minutes and resuspended after supernatant removal to reduce free viral particles. Thirty microliters of each sample was combined with UltraPure molecular grade water (Thermo, Inc.), 10X buffer (1x final concentration), dNTPs (0.1mM each final concentration), Cyano_SP_F primer (1.0 μM final concentration), Cyano_SP_R_519R fusion primers (R1 and R2 combined, 0.01 μM each final concentration), S-*-Univ-1100-a-A-15 16S reverse primer (1.0 μM final concentration), bovine serum albumin (0.02 mg/mL final concentration), Tween-20 (0.8% v/v final concentration), and Phusion High-Fidelity DNA Polymerase (1.5U; New England BioLabs, Inc.) to a final volume of 75 μL (see Table S6 for primer sequences and citations). PCR reactions were combined with 450 μL of 4% UMIL EM90 oil (4% UMIL EM90 oil, 0.05% TritonX-100 v/v in mineral oil; Universal Preserv-A-Chem, Inc.) and emulsified by vortexing at max speed (~2,700 rpm) for one minute. Emulsions were loaded as 50 μL aliquots and amplified with the following conditions: denaturation at 94°C for 3 minutes; 33 cycles of denaturation at 94°C for 10 seconds, annealing at 54°C for 30 seconds, and extension at 72°C for 45 seconds; and a final extension of 72°C for 5

643 minutes (C1000, BioRad Labs., Inc.). Samples were immediately removed upon completion of
644 amplification and stored at -20°C until the emulsion was broken as described above.
645
646 *Confirmation of blocking primer efficacy prior to nested PCR amplification*
647 Blocking primers are used during nested PCR to prevent the annealing and amplification
648 of unfused genes from the emulsion PCR[26]. Prior to enriching our fusion products via nested
649 PCR (detailed below), we tested the ability of the blocking primers to prevent amplification of
650 unfused products at relevant 16S and RNR gene concentrations. To simulate our nested PCR
651 conditions, RNR and 16S rRNA gene copy numbers were quantified in all epicPCR reactions
652 post-cleanup by qPCR as described above. Next, RNR genes were amplified from Rhode River
653 samples with primers Cyano_II_F and Cyano_II_R_519R primers (0.3 μM final concentration
654 each, Table S6). 16S rRNA genes were similarly amplified with primers 27F and 1492R (0.3 μM
655 final concentration each, Table S6). Unfused RNR and 16S rRNA gene amplicons were
656 combined at copy numbers equal to the highest observed across all samples after epicPCR as
657 determined by qPCR. Finally, the unfused RNR and 16S rRNA gene mixture was run through
658 nested PCR and barcoding PCR reactions with all environmental samples as described below.
659 Samples were run on a 1.5% agarose gel and visualized with SYBR Safe DNA gel stain
660 (Invitrogen, 1x final concentration). Fusion products were observed in the unfused mix control
661 when no blocking primers were used. However, with the addition of blocking primers no fusion
662 products were detected in the unfused mix control. We therefore proceeded with nested PCR
663 amplification of the Chesapeake Bay fusion products with blocking primers (see below).
664
665 *Enrichment of viral-host fused amplicons*

Fused amplicons were enriched by nested PCR. Three microliters of column-purified DNA were combined with UltraPure molecular grade water (Thermo, Inc.), 10X buffer (1x final concentration), dNTPs (0.1mM each final concentration), Cyano_SP_Nested_FA primer (0.3 μM final concentration), 16S reverse primer PE_16S_V4_E786_R (0.3 μM final concentration), forward and reverse blocking primers U519F-block10 and U519R-block10 (1.0 μM final concentration each; to ensure no amplification of unfused genes, see above), and Phusion High-Fidelity DNA Polymerase (0.5U; New England BioLabs, Inc.) to a final volume of 25 μL (see Table S6 for primer sequences and citations). Samples were amplified with the following conditions: denaturing at 94°C for 30 seconds; 30 cycles of denaturing at 94°C for 10 seconds, annealing at 54°C for 30 seconds, and extension at 72°C for 45 seconds; and a final extension of 72°C for 5 minutes. PCR reactions were cleaned by spin column purification (Zymo, Inc.) and DNA eluted in 20 uL of Tris-HCl. Samples were barcoded following enrichment and cleanup by PCR as described above with Cyano_SP_Nested_FB primer (0.3 μM final concentration, Table S6), reverse primer PE-IV-PCR-XXX with 8-mer barcodes (0.3 μM final concentration, Table S6), and forward and reverse blocking primers U519F-block10 and U519R-block10 (1.0 μM final concentration each, Table S6). Barcoded samples were run on a 1.5% agarose gel (Invitrogen) and visualized with SYBR Safe DNA gel stain (Invitrogen, 1x final concentration). Fusion bands were cut out and gel purified (Zymo, Inc.). Fusion products were quantitated on a Qubit 3.0 fluorometer (Invitrogen) and ten nanograms of DNA pooled per sample for sequencing.

Sequencing and quality filtering of viral-host fusion amplicons

Fused amplicon products were sequenced on a PacBio Sequel with Sequel v3 chemistry (University of Maryland Institute for Genome Sciences). Circular consensus sequences were obtained from raw reads with the following parameters: minimum signal-to-noise ratio (SNR): 3, minimum length: 500bp, minimum passes: 10, minimum read score: 0.75, minimum predicted accuracy: 0.90. Consensus sequences passing these thresholds were oriented by searching for primer sequences using Cutadapt[102] with an error rate of 0.01. Only sequences passing this error rate were retained. Following orientation, reads were demultiplexed in QIIME[88] with zero allowed barcode mismatches. Fused gene products were split by gene and primer sequences identified with Cutadapt[102]. Only those sequences without mismatches in any primer sites (error rate of 0.01) were retained for further analyses. Finally, amplicons were filtered by size to remove truncated or chimeric reads. Only fusion sequences with RNR gene amplicons between 583 and 596bp and 16S rRNA gene amplicons between 249 and 262bp were retained. These sequences were split into separate RNR and 16S genes and their primer sequences trimmed with Cutadapt[102].

Analyzing viral and bacterial diversity from epicPCR fusion amplicons

Viral RNR and bacterial 16S rRNA gene sequences passing all quality filtering were clustered at 100% nucleotide identity with CD-HIT[103] to identify viral and bacterial populations. Only viral-host pairs observed in three or more libraries out of the 58 libraries total were considered positive interactions. 16S rRNA gene representative sequences were classified with the Ribosomal Database Project classifier Release 11.5[104]. RNR representative sequences were queried against the NCBI nr database via BLASTn to identify top hits. RNR gene sequences were aligned with MAFFT[96] using the L-INS-I setting. Sequences were trimmed to the region

711 G1380 to A2079 in the *E. coli nrdA* (RNR) gene and a maximum likelihood tree with 100
712 bootstrap replicates was made using the Jukes-Cantor substitution model in Phylogeny.fr
713 (http://phylogeny.lirmm.fr)[105]. The three different clades (Fig. 4a) were characterized by five of
714 more sequences sharing >95% average nucleotide identity. Viral-host interaction matrices for
715 each week of the timeseries can be found in Supplemental File 1. Viral-host interaction networks
716 were visualized in Cytoscape. Chesapeake Bay RNR amplicon sequences identified as CSP-like
717 Actinophage by epicPCR were queried against 2,034 freshwater phage genomes by BLASTn
718 with an e-value cutoff of 0.001. None of the freshwater phage genomes shared significant
719 homology with the RNR amplicon sequences.
720
721 *Viral dilution incubation experiments*
722 One liter of surface water was collected in July 2019 from the mouth of the Rhode River
723 (Edgewater, MD) at the Smithsonian Environmental Research Center. 120mL was filtered
724 through two 0.2 μm PES membrane filter (Millipore, Inc.) and the viral filtrate was collected.
725 Half of the viral filtrate was autoclaved to eliminate active viruses. The remaining viral filtrate
726 was left unamended. Bacterial communities were resuspended off of the 0.2 μm filters in 1mL
727 autoclaved viral filtrate by shaking for 10 minutes at medium speed on a MoBio Vortex Genie 2.
728 Resuspensions from each filter were pooled to create a single resuspended sample. Resuspended
729 cells were combined 1:120 with either autoclaved (no viruses) or un-autoclaved (with viruses)
730 viral filtrate. Samples were divided into twelve 20 mL replicates. Six replicates without viruses
731 and six replicates with viruses were incubated at room temperature. 500 μL aliquots were taken
732 at the beginning of the incubation and approximately every 24 hours over three days. 16S rRNA
733 and 'Cyano SP'-like RNR gene counts were quantified at each time point by qPCR as described

734 above. 16S rRNA gene amplicon libraries were created as described above for the initial
735 community and the community after 75 hours for all replicates. Growth of individual bacterial
736 populations over the course of the incubation was calculated as the fold-change in abundance
737 from T0 to T75 using qPCR results and relative abundances. OTUs that displayed growth in at
738 least four treatment replicates (without viruses or with viruses) were assessed for significantly
739 (FDR $p < 0.05$) greater growth in one of the treatments using a Kruskal-Wallis test. Putative
740 susceptibility was calculated for OTUs with significantly greater growth in one of the treatments
741 as (OTU Abundance Fold Change Without viruses)/(OTU Abundance Fold Change With
742 Viruses). Higher values indicate greater presumed susceptibility to viral-induced mortality.
743
744 *Single-cell genomics*
745 Samples from May 17, 2017 were sent to the Bigelow Single Cell Genomics Center (East
746 Boothbay, ME; https://scgc.bigelow.org/) for sorting of the cyanobacteria and prokaryotic
747 fraction into 384-well plates. Sorting was conducted with the red fluorescence as a function of
748 forward scatter, used to gate for "cyanobacteria", and Syto-9 stained DNA as a function of
749 forward scatter, used to gate for "all prokaryotes". Single cell genome amplification (SAG
750 Generation 2 using WGA-X® amplification[106]) of sorted cells was conducted at the Bigelow labs
751 and returned to JHU for marker analysis. RNR and 16S rRNA marker genes were amplified from
752 2 μL of a 1:100 dilution of the amplified genomes according to the protocol above, without the
753 use of emulsions. Positive amplicons were purified and sequenced with the forward primers at
754 the Genetic Resource Core Facility at JHU with an Applied Biosystems 3730xl DNA Analyzer.
755

Life Sciences Reporting Summary. Further information on experimental design and reagents is available in the Life Sciences Reporting Summary.

Data availability. Sequences associated with 16S rRNA libraries from environmental samples and incubation experiments, bacterial and viral shotgun libraries, and fusion amplicons from epicPCR have been deposited in NCBI under BioProject accession number PRJNA599167. Water physicochemical measurements and qPCR data have been deposited in the BCO-DMO database under datasets 757405 and 821955. Datasets used in this analysis include Global Ocean Virome 2.0 (GOV 2.0), NCBI non-redundant nucleotide database (nr), Tampa Bay metagenomic libraries (BioProject Accession No. PRJNA28619, PRJNA47459, and PRJNA52403), and Damariscotta River Estuary, ME, USA (BioProject Accession No. PRJNA357591).

References

1 Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801 (2007).

2 Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature microbiology* **3**, 870 (2018).

3 Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F. & Gordon, J. I. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Reviews Microbiology* **10**, 607 (2012).

4 Suttle, C. A. The significance of viruses to mortality in aquatic microbial communities. *Microbial ecology* **28**, 237-243 (1994).

5 Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465-470, doi:10.1038/nature16942 (2016).

6 Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689, doi:10.1038/nature19366 (2016).

7 Winget, D. M. *et al.* Repeating patterns of virioplankton production within an estuarine ecosystem. *Proceedings of the National Academy of Sciences* **108**, 11506-11511 (2011).

8 Chen, X. W. *et al.* Tide driven microbial dynamics through virus-host interactions in the estuarine ecosystem. *Water Res* **160**, 118-129, doi:10.1016/j.watres.2019.05.051 (2019).

9 Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences* **108**, E288-E297 (2011).

10 Flores, C. O., Valverde, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *The ISME journal* **7**, 520-532 (2013).

11 Jover, L. F., Cortez, M. H. & Weitz, J. S. Mechanisms of multi-strain coexistence in host–phage systems with nested infection networks. *Journal of theoretical biology* **332**, 65-77 (2013).

12 Våge, S., Storesund, J. E. & Thingstad, T. F. Adding a cost of resistance description extends the ability of virus–host model to explain observed patterns in structure and function of pelagic microbial communities. *Environmental microbiology* **15**, 1842-1852 (2013).

13 Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *Fems Microbiology Reviews* **40**, 258-272, doi:ARTN fuv04810.1093/femsre/fuv048 (2016).

14 Gregory, A. C. *et al.* Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109-+, doi:10.1016/j.cell.2019.03.040 (2019).

15 Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific data* **5**, 170203 (2018).

16 Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature communications* **7**, 10613 (2016).

17 Hatfull, G. F. Dark matter of the biosphere: the amazing world of bacteriophage diversity. *Journal of virology* **89**, 8107-8110 (2015).

18 Middelboe, M., Chan, A. M. & Bertelsen, S. K. in *Manual of Aquatic Viral Ecology* (eds S. W. Wilhelm, M. G. Weinbauer, & C. A. Suttle) 118-133 (American Society of Limnology and Oceanography, 2010).

19 Deng, L. *et al.* Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513**, 242-+, doi:10.1038/nature13459 (2014).
20 Mosier-Boss, P. A. *et al.* Use of fluorescently labeled phage in the detection and identification of bacterial species. *Appl Spectrosc* **57**, 1138-1144, doi:Doi 10.1366/00037020360696008 (2003).
21 Allers, E. *et al.* Single-cell and population level viral infection dynamics revealed by phage FISH, a method to visualize intracellular and free viruses. *Environmental microbiology* **15**, 2306-2318 (2013).
22 Tadmor, A. D., Ottesen, E. A., Leadbetter, J. R. & Phillips, R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**, 58-62 (2011).
23 Bickhart, D. M. *et al.* Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol* **20**, 18, doi:10.1186/s13059-019-1760-x (2019).
24 Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the resistome and plasmidome to the microbiome. *Isme Journal* **13**, 2437-2446, doi:10.1038/s41396-019-0446-4 (2019).
25 Labonte, J. M. *et al.* Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *Isme Journal* **9**, 2386-2399, doi:10.1038/ismej.2015.48 (2015).
26 Spencer, S. J. *et al.* Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *The ISME journal* **10**, 427 (2016).
27 Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology* **37**, 632-639 (2019).
28 Bench, S. R. *et al.* Metagenomic characterization of Chesapeake Bay virioplankton. *Appl. Environ. Microbiol.* **73**, 7629-7641 (2007).
29 Kan, J., Evans, S. E., Chen, F. & Suzuki, M. T. Novel estuarine bacterioplankton in rRNA operon libraries from the Chesapeake Bay. *Aquatic Microbial Ecology* **51**, 55-66 (2008).
30 Chen, F. *et al.* Diverse and dynamic populations of cyanobacterial podoviruses in the Chesapeake Bay unveiled through DNA polymerase gene sequences. *Environmental Microbiology* **11**, 2884-2892, doi:10.1111/j.1462-2920.2009.02033.x (2009).
31 Kan, J., Suzuki, M. T., Wang, K., Evans, S. E. & Chen, F. High temporal but low spatial heterogeneity of bacterioplankton in the Chesapeake Bay. *Appl. Environ. Microbiol.* **73**, 6776-6789 (2007).
32 Nasko, D. J. *et al.* Family A DNA polymerase phylogeny uncovers diversity and replication gene organization in the virioplankton. *Frontiers in microbiology* **9**, 3053 (2018).
33 Sakowski, E. G. *et al.* Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proceedings of the National Academy of Sciences* **111**, 15786-15791 (2014).
34 Dwivedi, B., Xue, B., Lundin, D., Edwards, R. A. & Breitbart, M. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evolutionary Biology* **13**, 33, doi:10.1186/1471-2148-13-33 (2013).

35 Harrison, A. O., Moore, R. M., Polson, S. W. & Wommack, K. E. Reannotation of the ribonucleotide reductase in a cyanophage reveals life history strategies within the virioplankton. *Frontiers in microbiology* **10**, 134 (2019).
36 Suzek, B. E., Huang, H. Z., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288, doi:10.1093/bioinformatics/btm098 (2007).
37 Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature communications* **8**, 15892 (2017).
38 Kavagutti, V. S., Andrei, A. S., Mehrshad, M., Salcher, M. M. & Ghai, R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome* **7**, 135, doi:10.1186/s40168-019-0752-0 (2019).
39 Tzortziou, M. *et al.* Tidal marshes as a source of optically and chemically distinctive colored dissolved organic matter in the Chesapeake Bay. *Limnology and Oceanography* **53**, 148-159, doi:DOI 10.4319/lo.2008.53.1.0148 (2008).
40 Jordan, T. E., Pierce, J. W. & Correll, D. L. FLUX OF PARTICULATE MATTER IN THE TIDAL MARSHES AND SUBTIDAL SHALLOWS OF THE RHODE RIVER ESTUARY. *Estuaries* **9**, 310-319, doi:10.2307/1351410 (1986).
41 Chai, T. J. Characteristics of Escherichia-Coli Grown in Bay Water as Compared with Rich Medium. *Applied and Environmental Microbiology* **45**, 1316-1323, doi:Doi 10.1128/Aem.45.4.1316-1323.1983 (1983).
42 Martiny, J. B., Riemann, L., Marston, M. F. & Middelboe, M. Antagonistic coevolution of marine planktonic viruses and their hosts. (2014).
43 Sieradzki, E. T., Ignacio-Espinoza, J. C., Needham, D. M., Fichot, E. B. & Fuhrman, J. A. Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat Commun* **10**, 1169, doi:10.1038/s41467-019-09106-z (2019).
44 Moniruzzaman, M. *et al.* Virus-host relationships of marine single-celled eukaryotes resolved from metatranscriptomics. *Nat Commun* **8**, 16054, doi:10.1038/ncomms16054 (2017).
45 Duffy, S., Turner, P. E. & Burch, C. L. Pleiotropic costs of niche expansion in the RNA bacteriophage Φ6. *Genetics* **172**, 751-757 (2006).
46 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359, doi:10.1126/science.1261359 (2015).
47 Deng, L. *et al.* Contrasting life strategies of viruses that infect photo-and heterotrophic bacteria, as revealed by viral tagging. *MBio* **3**, e00373-00312 (2012).
48 Adriaenssens, E. M. & Cowan, D. A. Using Signature Genes as Tools To Assess Environmental Viral Ecology and Diversity. *Applied and Environmental Microbiology* **80**, 4470-4480, doi:10.1128/Aem.00878-14 (2014).
49 Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat Commun* **8**, 15892, doi:10.1038/ncomms15892 (2017).
50 Martinez-Hernandez, F. *et al.* Droplet Digital PCR for Estimating Absolute Abundances of Widespread Pelagibacter Viruses. *Front Microbiol* **10**, 1226, doi:10.3389/fmicb.2019.01226 (2019).
51 Vik, D. R. *et al.* Putative archaeal viruses from the mesopelagic ocean. *Peerj* **5**, e3428, doi:10.7717/peerj.3428 (2017).

52 Jover, L. F., Romberg, J. & Weitz, J. S. Inferring phage–bacteria infection networks from time-series data. *Royal Society Open Science* **3**, 160654 (2016).
53 Brankatschk, R., Bodenhausen, N., Zeyer, J. & Burgmann, H. Simple Absolute Quantification Method Correcting for Quantitative PCR Efficiency Variations for Microbial Community Samples. *Applied and Environmental Microbiology* **78**, 4481-4489, doi:10.1128/Aem.07878-11 (2012).
54 Baran, N., Goldin, S., Maidanik, I. & Lindell, D. Quantification of diverse virus populations in the environment using the polony method. *Nat Microbiol* **3**, doi:10.1038/s41564-017-0045-y (2018).
55 Russell, D. A. & Hatfull, G. F. PhagesDB: the actinobacteriophage database. *Bioinformatics* **33**, 784-786, doi:10.1093/bioinformatics/btw711 (2017).
56 Jensen, E. C. *et al.* Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. *Applied and Environmental Microbiology* **64**, 575-580 (1998).
57 Peters, D. L., Lynch, K. H., Stothard, P. & Dennis, J. J. The isolation and characterization of two Stenotrophomonas maltophilia bacteriophages capable of cross-taxonomic order infectivity. *Bmc Genomics* **16**, 664, doi:10.1186/s12864-015-1848-y (2015).
58 Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425-+, doi:10.1038/nature19094 (2016).
59 John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental microbiology reports* **3**, 195-202 (2011).
60 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
61 Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).
62 Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
63 Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
64 Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607 (2015).
65 Song, W.-Z. & Thomas, T. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics* **33**, 1873-1875 (2017).
66 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043-1055 (2015).
67 Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927, doi:10.1093/bioinformatics/btz848 (2019).
68 Schütze, T. *et al.* A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Analytical biochemistry* **410**, 155-157 (2011).
69 Bolyen, E. *et al.* QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. Report No. 2167-9843, (PeerJ Preprints, 2018).

70 DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069-5072 (2006).
71 Warwick-Dugdale, J. *et al.* Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *Peerj* **7**, e6800, doi:10.7717/peerj.6800 (2019).
72 Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology* **15**, 1428-1440 (2013).
73 Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*, <https://www.osti.gov/servlets/purl/1241166> (2014).
74 De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666-2669, doi:10.1093/bioinformatics/bty149 (2018).
75 Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* **37**, 540-+, doi:10.1038/s41587-019-0072-8 (2019).
76 Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *Plos One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
77 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).
78 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *Peerj* **3**, e985, doi:10.7717/peerj.985 (2015).
79 Ren, J. *et al.* Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, doi:10.1007/s40484-019-0187-4 (2020).
80 Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L. & Sullivan, M. B. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *Isme Journal* **11**, 7-14, doi:10.1038/ismej.2016.89 (2017).
81 Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498, doi:10.1126/science.1261498 (2015).
82 Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *Bmc Genomics* **17**, 930, doi:10.1186/s12864-016-3286-x (2016).
83 Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol* **37**, 29-37, doi:10.1038/nbt.4306 (2019).
84 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-U354, doi:10.1038/Nmeth.1923 (2012).
85 Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632-+, doi:10.1038/s41587-019-0100-8 (2019).
86 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *Bmc Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).
87 Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432, doi:10.1093/bioinformatics/btq675 (2011).

88 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335 (2010).
89 Warnes, G. R. *et al.* gplots: Various R programming tools for plotting data. (2015).
90 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012).
91 Jasna, V., Parvathi, A. & Dash, A. Genetic and functional diversity of double-stranded DNA viruses in a tropical monsoonal estuary, India. *Sci Rep-Uk* **8**, 16036, doi:10.1038/s41598-018-34332-8 (2018).
92 McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental Microbiology* **16**, 570-585, doi:Doi 10.1111/1462-2920.12184 (2014).
93 Allen, L. Z. *et al.* The Baltic Sea virome: Diversity and transcriptional activity of DNA and RNA Viruses. *Msystems* **2**, e00125-00116, doi:10.1128/mSystems.00125-16 (2017).
94 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic acids research* **38**, e132-e132 (2010).
95 Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282-1288 (2007).
96 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).
97 Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649 (2012).
98 Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113-3114 (2017).
99 Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11-16, doi:10.1093/nar/gkh152 (2004).
100 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:Doi 10.1006/Jmbi.1990.9999 (1990).
101 Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **35**, W52-W57, doi:10.1093/nar/gkm360 (2007).
102 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
103 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).
104 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261-5267, doi:Doi 10.1128/Aem.00062-07 (2007).
105 Dereeper, A. *et al.* Phylogeny. fr: robust phylogenetic analysis for the non-specialist. *Nucleic acids research* **36**, W465-W469 (2008).

106 Stepanauskas, R. *et al.* Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun* **8**, 10, doi:10.1038/s41467-017-00128-z (2017).

Correspondences and requests for materials should be addressed to Sarah Preheim (sprehei1@jhu.edu) and Eric Sakowski (e.g.sakowski@msmary.edu).

Acknowledgements

The authors would like to thank the Smithsonian Environmental Research Center and Katrine Lohan for providing access to their facilities during sample collection. This work was supported by the National Science Foundation Biological Oceanography (awards #1820652, #1829831 and #1756314) and a Gordon and Betty Moore Foundation Investigator Award (#3790). Part of this project was conducted using computational resources at the Maryland Advanced Research Computing Center (MARCC) and the Ohio Supercomputer Center (OSC) for High Performance Computing.

Author Contributions

EGS and SPP conceived of the work. EGS conducted all field-work, epicPCR analysis, and incubation experiments associated with this work. EGS, KAW, and SPP conducted the experimental and computational analysis for the bacterial metagenome and 16S rRNA gene libraries. EGS, FT, AAZ, and OZ conducted experimental and computational analysis for the viral metagenomic libraries. EGS, KAW, and SPP conducted the bioinformatic host prediction. EGS wrote the manuscript. EGS, AAZ, OZ, MBS, and SPP edited the manuscript.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to EGS and SPP.

Figure legends

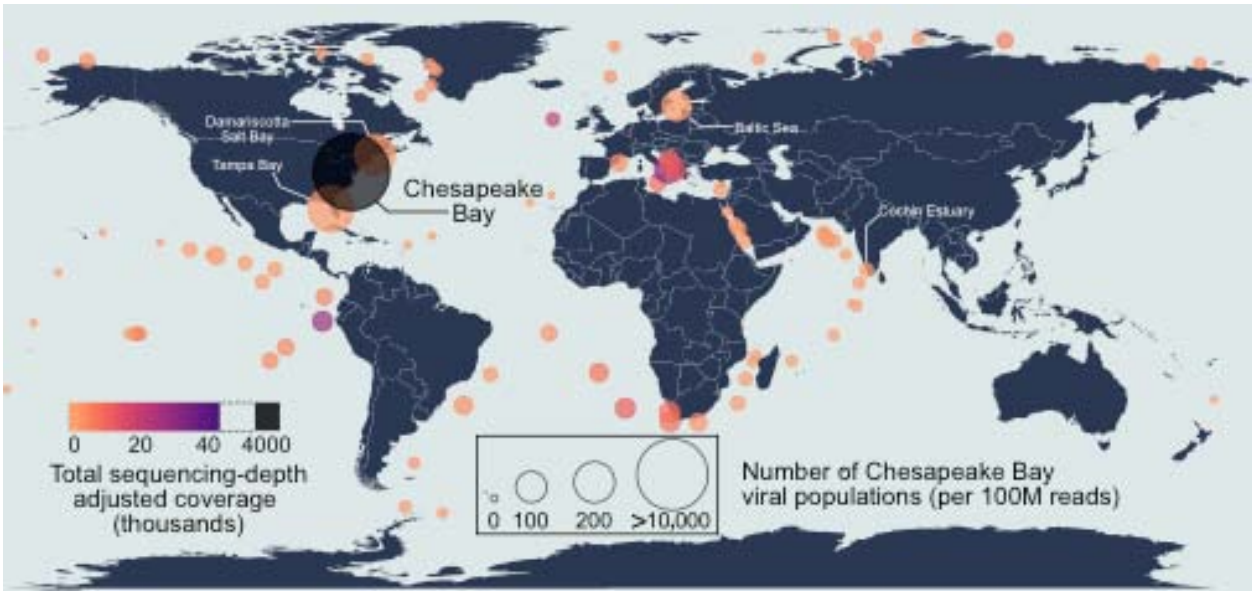

Figure 1. Global distribution and abundance of Chesapeake Bay viral populations. The Chesapeake Bay viral populations were highly endemic as less than 100 populations were shared with any other location in the global ocean viromes (GOV2.0) dataset (circle size) and recruited significantly less reads from the GOV2.0 stations (circle color). Similarly, when compared to other estuaries (labelled circles), less than 300 Chesapeake Bay viral populations were shared with these other estuaries with significantly less recruited reads from their libraries. A viral

population was considered detected in a non-Chesapeake Bay sample if the sample's reads covered at least 70% of the genomic length of the viral population. Due to differences in the sequencing depth between libraries, the number of shared viral populations and the sum of their coverage (i.e. total sequencing depth; see Materials and Methods) were adjusted for each library size. For the stations that have multiple samples, the maximum value is reported.

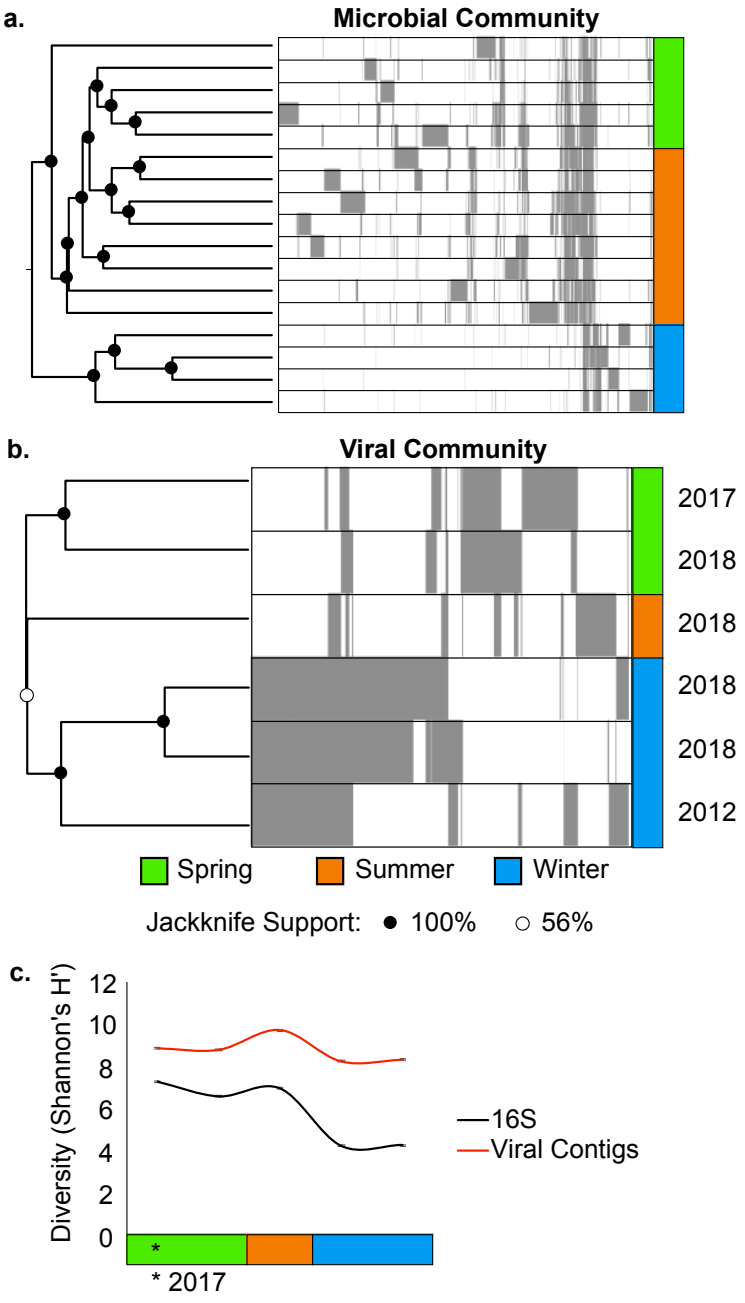

Figure 2. Seasonal dynamics of Chesapeake Bay bacterial and viral communities. Bacterial and viral surface water communities were collected from the mouth of the Rhode River, a sub-estuary of the Chesapeake Bay, in May 2017 and again in May, August, and December 2018. a.) Bray-Curtis hierarchical clustering of bacterial communities based on 16S rRNA gene sequences shows that microbial communities are most similar within the same season. Libraries were sub-

sampled at 150,000 observations with ten jackknifed replicates. Individual sequence variants are colored gray (present) or white (absent). b.) Bray-Curtis hierarchical clustering of viral communities from viral populations > 5kb collected across different years (see Materials and Methods) shows that seasonal community similarity exceeds within-year community similarity. Libraries were sub-sampled at 300,000 observations. Individual viral populations are colored gray. For the deeply-sequenced 2012 sample, only viral populations that were observed in the shallower 2017-2018 samples (4,328 of 12,736 viral populations, 34% of the 2012 populations) were included for clarity. c.) Bacterial and viral diversity (Shannon's H' index) for paired 16S rRNA gene and viral metagenomes were significantly correlated ($r_s = 0.9$, $p = 0.04$). Paired samples were from May 2017 (indicated by *), May 2018, August 2018, and December 2018 (n = 5 biologically independent samples). Viral metagenomes were subsampled to the smallest metagenome (15 M reads) prior to assembly for all Shannon's H' diversity analyses. Diversity indices represent the mean of jackknifed replicates (n = 10 subsamples of the same initial community) at a sampling depth of 10,000 observations. Error bars are SD.

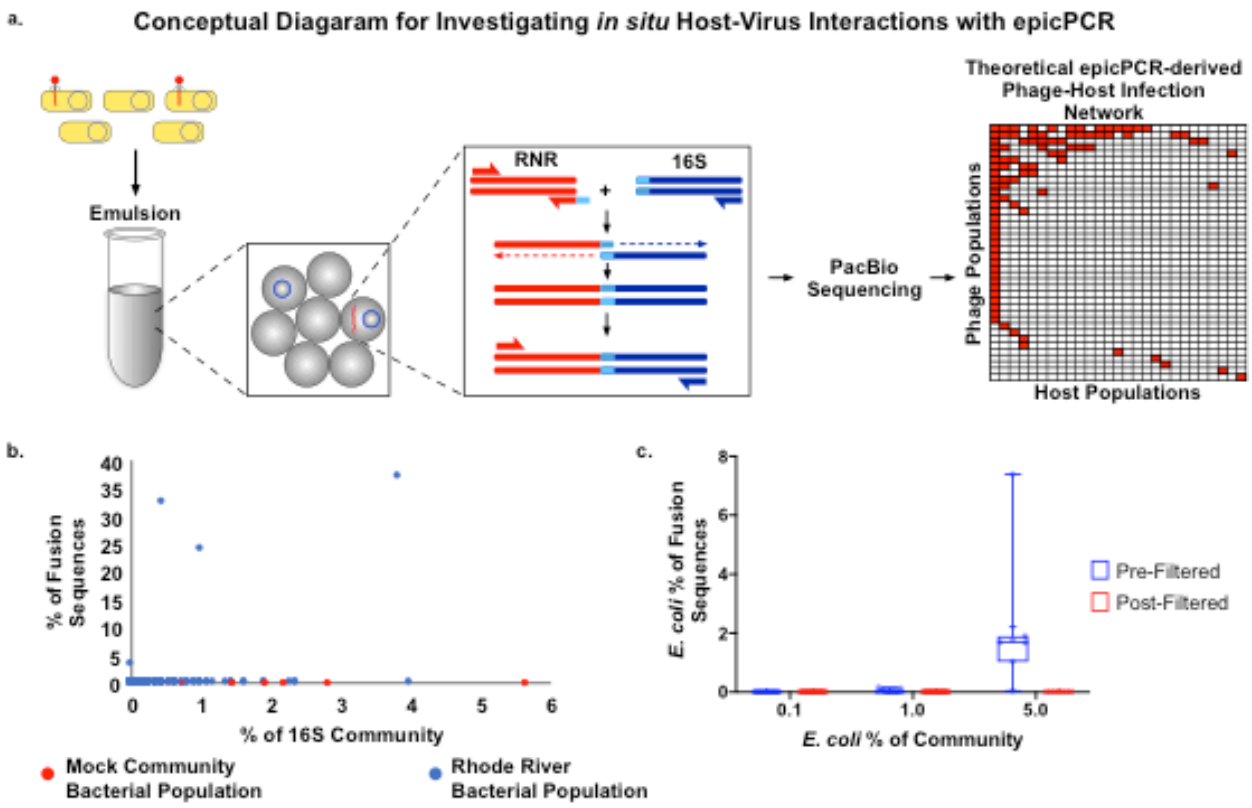

Figure 3. EpicPCR identifies phage-host interactions in the environment without cultivation. a.) Overview of the experimental design for epicPCR, identifying phage-host interactions through single-cell isolation in emulsion droplets and virus and host marker gene fusion. Left: individual cells are isolated within emulsion droplets and the genome of the host (blue circle) and virus (red curved lines) serve as the template for the gene fusion reaction. Middle: fusion PCR joins and amplifies viral and host marker genes from actively infected cells within emulsion droplets. RNR (ribonucleotide reductase; red) and 16S (16S rRNA gene; blue) are joined through an overlapping primer sequence (light blue). Right: fused amplicons are sequenced and analyzed to identify the network of viral-host interactions in the environment. b-c.) Control experiments to test epicPCR specificity. Specificity of the method was tested by spiking Chesapeake Bay water samples with a mock community of ten OTUs (cultivated cells, inactivated) or a single uninfected host (E. coli) prior to emulsification. b.) Proportion of fusion sequences belonging to uninfected mock (red) and Chesapeake Bay (blue) community members as a function of their

community abundance. Uninfected mock community sequences were not found to be associated within any fusion products. c.) Proportion of fusion products containing E. coli when spiked into Chesapeake Bay samples at 0.1% (n = 17), 1% (n = 17), and 5% (n = 7) of the community. Post-filtered sequences are those interactions observed in a minimum of three libraries. Box and whisker markers represent the minimum, first quartile, median, third quartile, and maximum values.

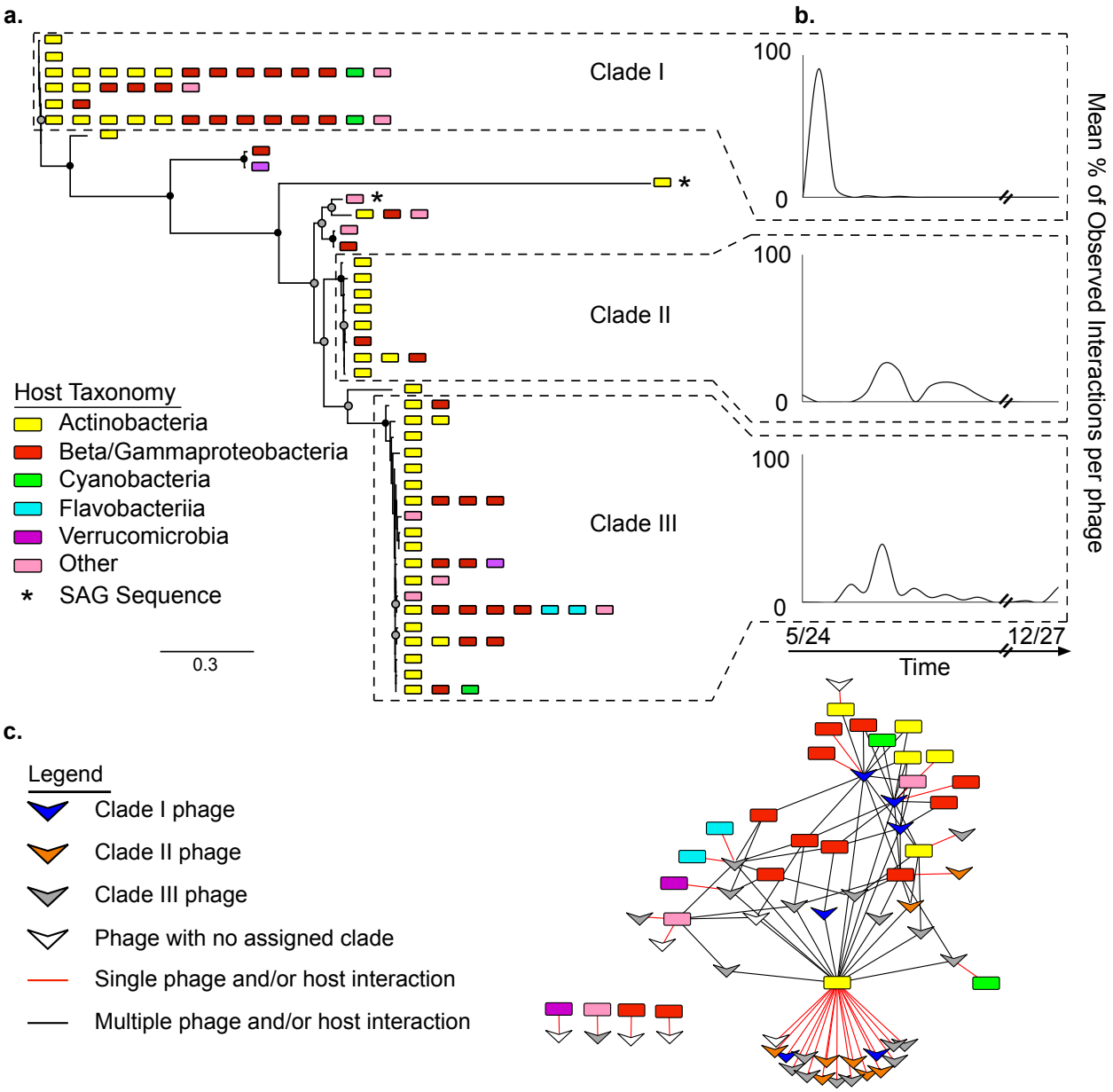

Figure 4. Abundance, diversity, and ecology of CSP-like Actinophage virus-host interactions as determined by epicPCR in the Chesapeake Bay from May to December 2018. a.) Maximum likelihood tree with 100 bootstrap replicates of Chesapeake Bay CSP-like Actinophage RNR gene sequences (G1380 to A2079 in *E. coli nrdA*, See Materials and Methods). Nodes are colored by bootstrap support: Black > 75%; Grey 50 – 75%; no color < 50%. Only RNR gene sequences that were observed with the same host (100% nucleotide identity of 16S rRNA gene) in at least three epicPCR fusion amplicon libraries were included. Phage RNR gene sequences

formed three main clades where five or more RNR sequences shared > 95% nucleotide identity. Phage RNR sequences were clustered at 100% nucleotide identity, resulting in 40 unique sequences shown on the tree. Host taxonomy identified by 16S rRNA gene homology for each phage RNR gene sequence is indicated. In some cases, closely related viruses were associated with different hosts. Scale bar represents nucleotide substitutions per site. b.) Mean percentage of total identified interactions observed at each sample time point by phage RNR clade. Phage-host interactions were investigated by epicPCR weekly from May to August 2018 and again weekly in December 2018. Clade I phage RNR sequences were primarily observed in late spring, while RNR sequences from Clades II and III were observed throughout the summer. c.) Aggregate network of Chesapeake Bay CSP-like Actinophage-host interactions from May to December 2018. Host taxonomy is colored as in panel A and is represented with rectangular shapes. Phage taxonomy is represented by the color map on the left and is represented by v-shapes. Lines represent virus-host interactions observed with epicPCR, colored according to viral host range. Most CSP-like Actinophage RNR sequences were associated with a single Actinobacteria host 16S rRNA gene identified as Luna-1 member *Rhodoluna*.

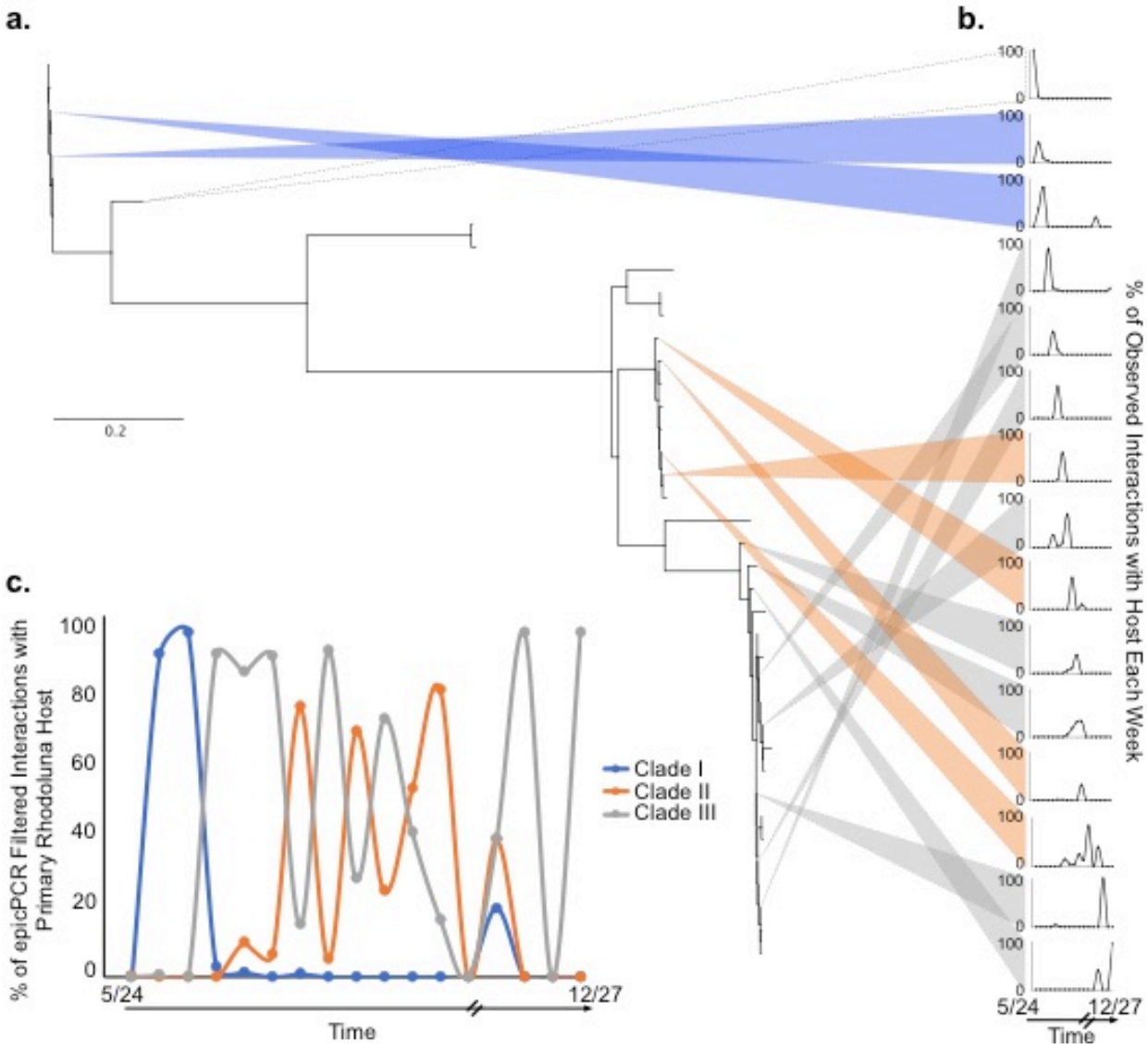

Figure 5. Phage interactions with Rhodoluna host observed in the Chesapeake Bay from May to August 2018 and again in December 2018 as revealed by epicPCR. In total, 31 of 40 phage populations identified by epicPCR interacted with this host. a.) Maximum likelihood tree with 100 bootstrap replicates of Chesapeake Bay phage ribonucleotide reductase (RNR) genes (G1380 to A2079 in E. coli nrdA, See Materials and Methods). RNR gene sequences were clustered at 100% nucleotide identity prior to phylogenetic analysis. Phage RNR gene sequences formed three main clades where RNR sequences shared > 95% nucleotide identity. Scale bar represents nucleotide substitutions per site. b.) Proportion of total observed interactions

identified by epicPCR each week between Rhodoluna host and each identified phage RNR gene sequence. Colored wedges link phage populations to their corresponding dynamics over time. Wedge colors correspond to phage clades: blue: Clade I; orange: Clade II; grey: Clade III; no color: no assigned clade. Only the most abundant phage interactions were depicted for simplicity. c.) Proportion of phage interactions by clade with Rhodoluna host over time identified by epicPCR. Blue: Clade I; orange: Clade II; grey: Clade III.

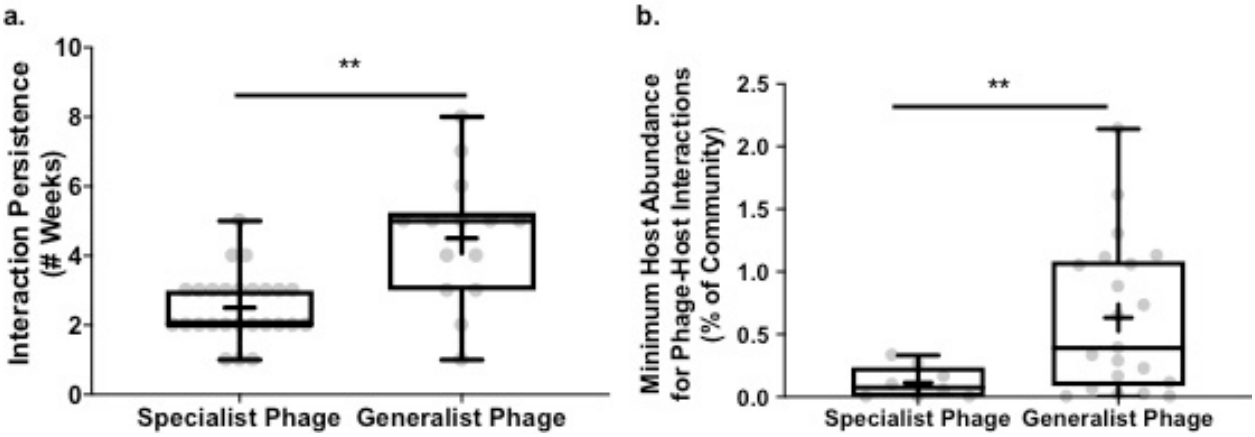

Figure 6. Biological trade-offs and ecological patterns related to viral and bacterial interactions revealed by epicPCR. a.) Total interaction persistence observed by epicPCR for CSP-like Actinophage RNR sequences between May and December 2018 in the Chesapeake Bay. RNR sequences associated with a single host in the epicPCR amplicons were designated specialist phages (n = 26), while RNR sequences associated with multiple hosts in the epicPCR amplicons were designated generalist phages (n = 14) (** = two-tailed T Test p = 0.002 for normally distributed data). Normal distributions were determined by the D'Agostino and Pearson test (p > 0.05). Box and whisker markers represent the minimum, first quartile, median, third quartile, and maximum values. The mean is indicated by "+". b.) Minimum host abundance (16S rRNA gene relative abundance) of interactions with specialist phage RNR gene sequences (n = 8) and generalist phage RNR gene sequences (n = 21) observed in epicPCR fusion amplicons from samples collected between May and December 2018 in the Chesapeake Bay (** = two-tailed T Test p = 0.001 for normally distributed data). Minimum host relative abundances indicate the lowest observed host abundance across timepoints where the host was associated by epicPCR with a generalist or specialist phage. All hosts identified by epicPCR were present in at least three 16S rRNA gene amplicon libraries. Data points where host abundance equals zero (n =

5/29 observations) indicate an observed phage-host interaction in epicPCR amplicons but no host detected in the 16S library for that time point. Omitting instances where the host was undetected in 16S rRNA gene amplicon libraries or using the next lowest abundance observation had no impact on the result. Normal distributions were determined by the D'Agostino and Pearson test ($p > 0.05$). Box and whisker markers represent the minimum, first quartile, median, third quartile, and maximum values. The mean is indicated by "+".