










Long-read powered viral metagenomics in the oligotrophic Sargasso Sea

Received: 30 March 2023

Accepted: 24 April 2024

Published online: 14 May 2024

 Check for updates

Joanna Warwick-Dugdale ^{1,2,9}✉, Funing Tian ^{3,9}, Michelle L. Michelsen¹, Dylan R. Cronin^{3,4}, Karen Moore ¹, Audrey Farbos¹, Lauren Chittick³, Ashley Bell ¹, Ahmed A. Zayed ^{3,4}, Holger H. Buchholz^{1,5}, Luis M. Bolanos ¹, Rachel J. Parsons^{6,7}, Michael J. Allen ¹, Matthew B. Sullivan ^{3,4,8} & Ben Temperton ¹✉

Dominant microorganisms of the Sargasso Sea are key drivers of the global carbon cycle. However, associated viruses that shape microbial community structure and function are not well characterised. Here, we combined short and long read sequencing to survey Sargasso Sea phage communities in virus- and cellular fractions at viral maximum (80 m) and mesopelagic (200 m) depths. We identified 2,301 Sargasso Sea phage populations from 186 genera. Over half of the phage populations identified here lacked representation in global ocean viral metagenomes, whilst 177 of the 186 identified genera lacked representation in genomic databases of phage isolates. Viral fraction and cell-associated viral communities were decoupled, indicating viral turnover occurred across periods longer than the sampling period of three days. Inclusion of long-read data was critical for capturing the breadth of viral diversity. Phage isolates that infect the dominant bacterial taxa *Prochlorococcus* and *Pelagibacter*, usually regarded as cosmopolitan and abundant, were poorly represented.

Bacteriophages are major drivers of both biogeochemical cycles and fitness selection of ecotypes. Ocean viruses directly influence availability of carbon via host lysis¹; structure microbial communities through negative density dependent selection²; and alter biochemical function through co-evolution^{3–5} and metabolic hijacking/reprogramming^{6,7} (reviewed in^{8,9}). Global ocean datasets characterising microbial communities have enabled machine learning and ecosystem modelling approaches to identify which of the many marine microbiota best predict key ecosystem features, including identifying viruses as the best predictor of carbon flux from the surface to deep oceans¹⁰. In the Sargasso Sea, abundance of virus-like particles has seasonal and depth-associated structure with a maximum concentration observed at 80 m.

Viral abundance correlates positively to abundance of the dominant phototrophs, *Prochlorococcus*, and negatively to abundance of the dominant heterotrophs, SARI1¹¹. Curiously, pelagiphages (phages that infect SARI1) have been reported as globally ubiquitous and abundant^{12–15}, but do not contribute significantly to the variance in virus-associated carbon export to depth in the oligotrophic ocean, which is primarily driven by phages infecting *Synechococcus*¹⁰. This appears to raise a paradox: Pelagiphages dominate global oceans in abundance, yet appear insignificant in both driving carbon export and restructuring cellular communities in the Sargasso Sea?

One hypothesis of the disconnect between host turnover and viral abundance in SARI1 is that chronic infection is more prevalent than

¹School of Biosciences, University of Exeter, Exeter, Devon EX4 4SB, UK. ²Plymouth Marine Laboratory, Plymouth, Devon PL1 3DH, UK. ³Center of Microbiome Science and Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. ⁴EMERGE Biology Integration Institute, Ohio State University, Columbus, OH 43210, USA. ⁵Department of Microbiology, Oregon State University, Corvallis, OR 97331, USA. ⁶Bermuda Institute of Ocean Sciences, St. George's, GE 01, Bermuda. ⁷School of Ocean Futures, Arizona State University, Tempe, AZ, US. ⁸Department of Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA. ⁹These authors contributed equally: Joanna Warwick-Dugdale, Funing Tian.

✉ e-mail: jo.warwick@gmail.com; b.temperton@exeter.ac.uk

virulent lysis in SAR11 host-virus dynamics, as suggested by low transcriptional activity of pelagiphages in a temperate coastal system¹⁶. Alternatively, due to the enormity of SAR11 populations, a small proportion of susceptible cells within a much larger population of resistant cells could sustain large pelagiphage populations, as long as susceptible hosts possessed some ecological advantage over resistant conspecifics, as observed in *Synechococcus*¹⁷. More recently, observations of very low levels of *Prochlorococcus* infection (0.35–1.6%) in oligotrophic waters with high cyanophage abundance has been hypothesised to result from a combination of host resistance, low phage adsorption rates and rapid loss of infectivity of virions¹⁸. To date, our understanding of the structure of the viral communities in the Sargasso Sea have been limited to seasonal patterns in abundances of viral-like particles¹¹ and genomic analysis of isolated phages^{14,19–22}.

Here, we used metagenomics to characterise the viral communities at 80m and 200m in both the cellular and viral fractions of the stratified Sargasso Sea during a four-day cruise in July 2017. Long read viral fraction sequencing was used to overcome assembly fragmentation due to microdiversity and to improve recovery of virally encoded hypervariable regions (HVRs) to facilitate evaluation of their role in niche-adaptation^{23–26}. We compared viral abundance between paired cellular and viral fractions to show that the composition of the viral fraction population did not reflect that of the associated cellular fraction. Viral communities were distinct between depths and comprised many viral populations that were undetected in previous global ocean viral surveys. Phages known to infect SAR11 and *Prochlorococcus*, from both isolates and metagenomic viral contigs where host could be determined, were poorly represented in the viral fractions, hinting at potential low viral contribution to cellular turnover and nutrient recycling in these taxa during the sampling campaign.

Results

Overview

In marine microbial communities, replication is known to be linked to diel cycles, as observed in both photosynthetic- and heterotrophic bacteria, and picoeukaryotes^{27,28}. Assuming that diel cycles would also influence phage production, we sampled cellular and viral fractions for metagenomics over three diel periods to maximise diversity of recovered phage genomes from the Sargasso Sea. DNA was collected from both the cellular fraction ($> 0.2 \mu\text{m}$; $n = 12$; included host DNA, lysogenic viruses, actively replicating viruses, and any free viruses attached to cells; $0.22 \mu\text{m}$ filtration and phenol-chloroform extraction method), and the “viral fraction” ($< 0.2 \mu\text{m}$, $n = 12$; includes virus-like particles; ferric chloride flocculation and spin column extraction method). Samples were taken at depths of 80 m (the ‘viral particle maximum’¹¹), and 200 m (the mesopelagic), to maximise the diversity of viruses surveyed. DNA recovered from all samples was sequenced to a depth of 12.1 Gbp on the Illumina platform to generate short reads. Additionally, viral-fraction samples from 80 m were sequenced via Nanopore to generate long reads ($n = 3$, 14.7 Gbp total). Short reads were assembled alone and used for hybrid long- and short-read assembly (VirION²³) before identification of viral contigs via VirSorter²⁹ (all bioinformatic processes are described in detail in Methods and Materials and summarised in Supplementary Fig. 1).

Results and discussion

Inclusion of long reads improves virus population recovery

In total, 3514 putative viral contigs $> 10 \text{ kb}$ in length were recovered from the Sargasso Sea; 2049 of these were derived from short-read assemblies (12 cellular fraction and 12 viral fraction samples), and a further 1465 were generated by assembly of VirION reads (three viral fraction samples). Contigs were clustered into 2301 viral populations²⁴ (equivalent to species). In 1410 (61%) of the viral populations the longest contig (selected as the cluster representative) was derived from long-read sequencing of pooled 80 m samples. Only 163 (7%) of

the viral populations contained both long- and short-read derived contigs. No population clusters with two or more members were comprised exclusively of long-read contigs, suggesting either the coverage of the virome in long read data was low, or that short-read sequencing captured genomes across all viral populations, but such genomes were fragmented and long-read assembly provided longer representatives. When the 24 short-read assemblies were processed without long reads, 1044 viral populations were identified, indicating that 55% (1257) of the total viral populations reported here were only captured with the inclusion of long-read sequencing.

An important question is whether viral populations represented by long-read assembled contigs were successfully recovering contigs from short-read assemblies that had fragmented below the 10kb cut-off, or whether the additional sequencing depth offered by long-reads enabled better recovery of ‘rare’ viruses. If the former is true, fragmented contigs from abundant viruses should align to the genomes of their respective long-read representative. If the latter is true, clusters with long-read representatives should be enriched at low relative abundance values – for viral populations with high relative abundance, there would be sufficient data to recover the genome without the addition of long reads. Mapping of short-read data to viral population representatives indicated viral populations with long-read representatives were abundant in the Sargasso Sea (Fig. 1A). Alignment of short-read population cluster members to long-read population cluster representatives illustrated fragmentation and/or poor overall recovery of 115 viral population genomes across a range of relative abundances in the short-read assemblies (Supplementary Fig. 2A). If assembly of viral genomes from short read data is restricted by genome coverage, one would expect that the sum of fragment lengths from abundant viruses would be similar to that of the long-read representative of their cluster, whereas low abundance viruses would cover less of the long-read representative genome. However, the extent to which genomes from long read assemblies were recovered by mapped short-read viral contigs did not correlate to the relative abundance of those contigs, when either viral contigs $> 10 \text{ kb}$ were used (Supplementary Fig. 2B), or when short-read contigs $> 1 \text{ kb}$ were used (Supplementary Fig. 2C) suggesting assembly breakages were not a result of low coverage. Representative contigs from long-read assemblies were 38.5% longer than those from short-read assemblies (at median lengths of 17,372 bp and 15,755 bp, respectively; two-sided Mann-Whitney U test, $p < 0.001$) (Fig. 1B). Inclusion of long-read data was also critical for enabling the recovery of hypervariable regions in viral genomes predicted to encode proteins involved in host recognition, DNA synthesis and DNA packaging (Supplementary Table 1; Supplementary Discussion). Together, these results suggest that long-read sequencing of viromes enhanced the capture of both different and more complete viral genomes from the Sargasso Sea compared to short-read technology alone.

Sargasso Sea Viruses formed a distinct community

We postulated whether Sargasso Sea viruses were endemic or globally distributed. We calculated the distribution of Sargasso Sea viruses from global oceanic viral metagenomes (GOV 2.0³⁰), and found that 800 (45.7%) of the viral genomes from this study were represented within other subsampled (to 5 million reads) global oceanic viral metagenomes (GOV 2.0) (Fig. 2). The log odds of finding a virus from the Sargasso Sea represented in a GOV 2.0 sample from a temperate/tropical epipelagic sample was negatively correlated to the availability of nitrate/nitrite and phosphate (logistic regression, nitrite/nitrate log odds = -0.13208 ; phosphate log odds = -0.13208 , $p < 0.05$ for both), suggesting greater frequency in samples from similarly warm oligotrophic oceanic regions (e.g., TARA_R10000455; Fig. 2). Sargasso Sea viruses were absent from polar regions, consistent with previously reported ecological patterns of ocean viral communities³⁰. However, 951 (54.3%) of the Sargasso Sea viral populations were not represented

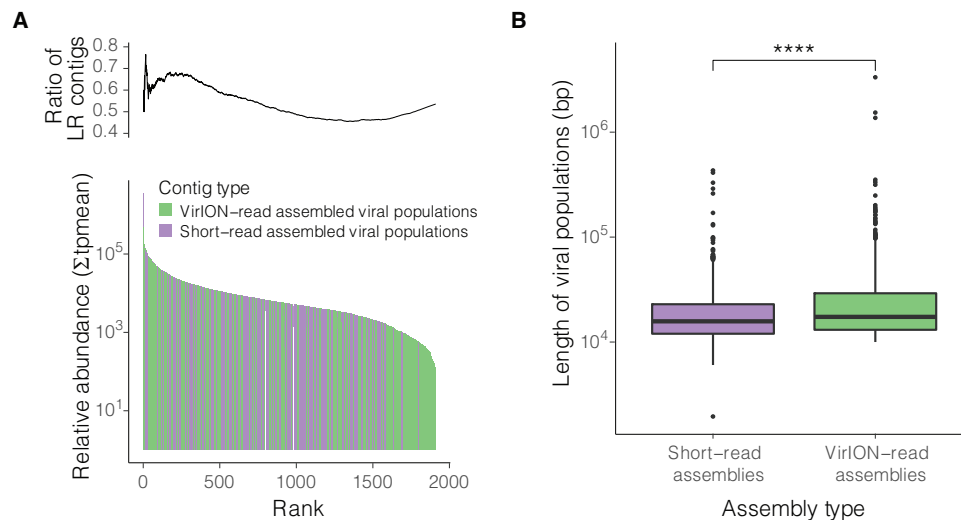


Fig. 1 | BATS viral populations. **A** Histogram of sequencing-depth adjusted coverage for viral populations ($n = 2301$). Long-read sequencing was able to rescue more viral populations, and these viral populations were abundant (Σ tpmean: sum of mean relative abundance of viral populations excluding < 5 th and > 95 th percentile). Inset: Cumulative ratio of viral populations derived from long reads (LR). **B** Boxplots showing statistically significant differences in lengths of viral

representative contigs assembled from short reads ($n = 891$) and VirION reads ($n = 1410$) (median lengths 15,755 bp and 17,372 bp, respectively; two-sided Mann-Whitney U test, **** $p = 8.08 \times 10^{-9}$). Boxes represent upper and lower quartiles; whiskers represent 1.5x interquartile range; individual points represent outliers; Centre line represents median population length.

in the global ocean dataset, suggesting that these viruses are either endemic to this subtropical region, or are enriched at this site and below the detection limit elsewhere. Twenty-four of these ‘endemic/enriched’ viruses were among the top 100 most abundant viruses at this site, indicating their prevalence in the Sargasso Sea. In contrast, our previous application of long-read sequencing to coastal Western English Channel viruses revealed that viral contigs obtained from a single sample were abundant in global marine viromes²³.

Inter-community diversity of the Sargasso Sea viral populations was compared to previously established patterns in GOV2.0³⁰, and revealed that Sargasso Sea viral communities have a distinct community structure (analysis of variance: Adonis F-test; p -value = 0.001; Fig. 3). Bathypelagic GOV2.0³⁰ samples were very divergent (Supplementary Fig. 3A, B), supporting previous evidence that viral populations from deep samples are distinct from others³⁰. To improve resolution bathypelagic samples were removed (Fig. 3), as were three GOV 2.0 viromes (station 155_SUR, station 72_MES, station 102_MES) that were outliers in both PCoA and Shannon’ H analyses³⁰; Supplementary Fig. 3B). Sargasso Sea viral populations (sampling depth: 80 m and 200 m) were most similar to viromes from other temperate-tropical mesopelagic regions (Fig. 3; two-sided pairwise Adonis F-tests of Euclidean distance between centroids (viral fraction populations vs TT-MES: 0.155; cellular fraction populations vs TT-MES: 0.101); p -value = 0.001; Supplementary Dataset 1). A Similarity Percentages (SIMPER) analysis was performed to determine the key Sargasso Sea viruses contributing to the dissimilarity of this group from viruses from the other ecological zones sampled in the Global Ocean Virome (GOV2) dataset (Fig. 4; Supplementary Dataset 2). SIMPER analysis showed that 754 viruses captured in the long-read data explained 9.5% of the variance that discriminated Sargasso Sea viromes from the other viral communities/zones. Viral populations were classified into two sets (Set A and Set B) depending on their recruitment of viral reads from other GOV2 samples. A set of 80 viral populations from the Sargasso Sea (Set A) recruited reads from other global viromes and were important in discriminating between temperate and tropical epipelagic (TT-EPI), and temperate and tropical mesopelagic (TT-MES) populations, and between TT-EPI and arctic (ARC) populations. Within these viral populations, 79 out of 80 were comprised of singleton viral populations from long-read assembly that lacked even fragmented

short-read assemblies. The median number of Global Ocean Virome (GOV2) samples in which a phage in Set A was observed was 56 (52.5–58 95% CI; out of 142 total). Therefore, we hypothesise that these viruses are common across oceanic regions but missed from existing short-read viral metagenomic datasets. Five of these viral populations had greater global relative abundance than pelagic phage HTVC010P and 51 were more ubiquitous than HTVC010P and other isolates infecting SARI1 and *Prochlorococcus* spp., at middle and lower latitudes (Fig. 4). These observations suggest that the VirION approach captured globally distributed and ubiquitous viruses that would otherwise have remained unidentified. The remaining 675 viruses (Set B), were far less ubiquitous in global oceans, identified in a median of 10 (4.5–17 95% CI) out of 145 GOV2 samples. Four members from Set B recruited a large number of reads from at least one site in either the TT-MES or Antarctic biomes (Supplementary Dataset 2), implying some degree of viral import from either upwelling or ocean currents into the 80m and 200m samples. Overall, these results show that the viral community of the Sargasso Sea was distinct in the global ocean, for at least the duration of the 3-day sampling campaign, and supports the idea that ubiquitous viruses may contribute to the regionalisation of viral communities through their relative contribution to overall community structure³¹.

Viral communities were decoupled by fraction

The physical stratification of the ocean due to seasonal warming (and therefore reduced density) of the upper layers, combined with the attenuation of light with depth, has long been understood as an important factor in structuring pelagic, microbial communities down the water column^{32–35}. Likewise, the viral community composition in the Sargasso Sea differed significantly along the two ecological zones, as delineated by depth (Fig. 5A). These results support previous evidence that phages are vertically distributed³⁰ in the same way as their bacterial counterparts, corresponding to the stratification of the water column during summer³⁶. Further investigation comparing these results to viral communities sampled after the spring bloom would ascertain whether the same pattern is repeated throughout the year, when vertical stratification is either absent or minimised.

Here, we report significant differences in composition and membership of viruses between cellular and viral fractions from the same

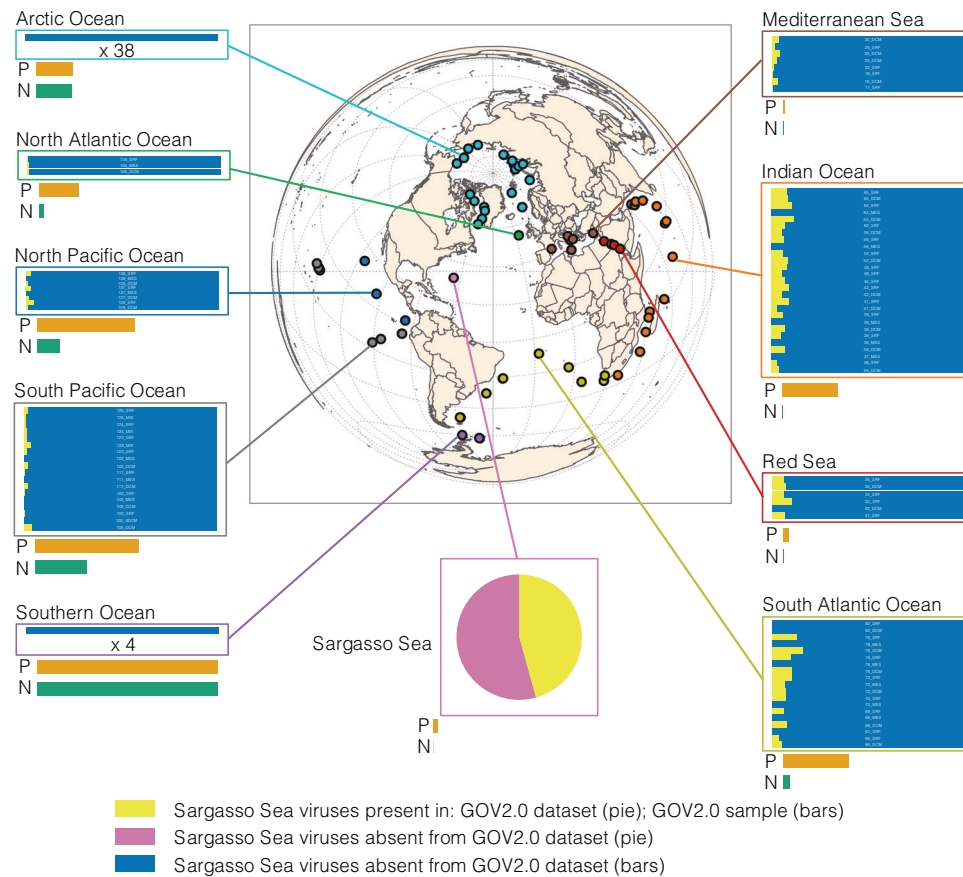


Fig. 2 | Global distribution of BATS viral populations. Dereplication of the combined Global Ocean Virome 2³⁰ (GOV 2.0) and Sargasso Sea dataset resulted in 152,979 viral populations (circular contigs or those with lengths ≥ 10 kb). Presence/absence of Sargasso Sea viruses are shown in various oceanic regions (within associated boxes). Under half of Sargasso Sea viruses (45.7%; coloured yellow at all sites) were observed as present elsewhere in the world's oceans. These viral sequences were identified as present in the GOV 2.0 dataset via competitive read recruitment of subsampled (5 million per sample) reads (mapped at $\geq 90\%$ read

length at $\geq 95\%$ identity) with $> 70\%$ genome coverage. The majority of Sargasso Sea viruses were not observed at GOV 2.0 sites (54.3%; coloured pink at the Sargasso Sea site; coloured blue at other sites), and appear endemic to the Sargasso Sea, (i.e., below the level of detection in the subsampled GOV 2.0 dataset). Levels of nitrogen ($\text{NO}_3 = \text{Nitrate} + \text{Nitrite}^{-1}$ ($\mu\text{mol/kg}$); N; coloured green) and Phosphate ($\text{PO}_4 = \text{Phosphate}^{-1}$ ($\mu\text{mol/kg}$); P; coloured orange) were normalised between sample sites (e.g., lowest site median: $P = 0$; highest site median: $P = 1$) and show that more Sargasso Sea virus tend to be present at oligotrophic GOV 2.0 sites.

depth. Bacteriophage populations in the Sargasso Sea formed discrete cellular fraction and viral fraction populations, at both the viral maximum depth (i.e., 80 m^{ll}) and in the mesopelagic (Fig. 5A). This zonal partitioning was also observed when long-read derived contigs were excluded from the analysis (Supplementary Fig. 4), indicating that it is not an artefact of sequencing technology. Sixty-five percent of viral populations from the short reads (678) were only detected in the 'viral fraction', whereas 7% (74) were discovered solely in the cell-associated fraction, and 28% (292) were detected in both fractions (Fig. 5B). Previous investigations of soil microbial communities and their phages have examined coupled cellular and viral fractions of samples for viral genomes and found that the majority (77%) were present in the cellular fraction, despite the samples for each fraction being collected years apart^{37,38}. However, when viromes and cellular fractions of the same samples were compared, as undertaken here, just 9% of viral populations were shared between fractions, with $> 90\%$ present in only the viral fraction, and $< 1\%$ unique to the cellular fraction³⁹. Because the Sargasso Sea cellular fraction here may include any free viruses attached to- or caught on cells during the filtration step of our protocol, it is not possible to discount the presence of DNA from free viruses here. The large surface area of filters used to separate cellular and viral fractions, and the relatively low concentration of bacterioplankton cells during the sampling campaign (80 m : $-5\text{--}7.7 \times 10^8 \text{ L}^{-1}$; 200 m : $-1.2\text{--}3.2 \times 10^8 \text{ L}^{-1}$) make it unlikely that viruses were removed

from the 'viral' fraction by clogging on filters. Thus, the effect of filtration is unlikely to cause the degree of dissimilarity between the free-particle and cell-associated fractions observed in our data. Presence-absence analyses showed that this decoupling of fractions was less pronounced in the 100 most abundant Sargasso Sea viruses (Supplementary Fig. 5), thus rare viruses may be partially driving the dissimilarity in viral- and cellular fraction viral populations. However, Bray-Curtis analysis is robust to the effects of sample size⁴⁰, so the partitioning effect illustrated via Principal Coordinates Analysis is not likely due to under-sampling of rare viruses.

One explanation for the decoupling of cell-associated and viral fraction populations is that the viral fraction population represents the integral of infections in the cellular fraction over time, whereas the cell-associated populations are a mix of prophages, remnant phages, active lytic infections and a small number of phage particles trapped on filters. Viral turnover in oligotrophic waters has been estimated at 2.2 days, compared to 0.82–1.3 days in coastal waters⁴¹. Here, sampling was conducted over three days to capture viruses across full host growth and lytic cycles to avoid biases inherent to snap-shot surveys, where asynchrony in infection cycles can cause apparent dissimilarities between free and cell-associated viral populations. Decoupling between free and cell-associated viral populations suggests that, at the time of sampling, viral turnover in the Sargasso Sea occurred over periods longer than our sampling campaign. Loss of infectivity and

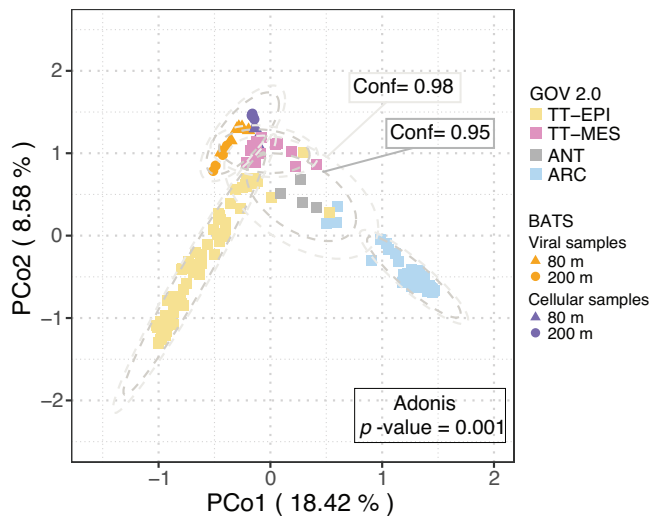


Fig. 3 | Inter-community diversity of Sargasso Sea and Global Ocean Virome 2 (GOV 2.0) viral populations. Principal coordinates analysis (PCoA) of a Bray-Curtis dissimilarity matrix calculated from mapping Global Ocean Virome 2³⁰ (GOV 2.0) reads and Sargasso Sea short reads to a combined dataset of Sargasso Sea and GOV 2.0 viral populations (contig lengths ≥ 10 kbp; $n = 152,979$); Viral community structure was suggested by ellipses drawn at 95% (inner) and 98% (outer) confidence intervals and analysis of variance (two-sided Adonis, p -value = 0.001); ARC Arctic, ANT Antarctic, TT-EPI Temperate and tropical epipelagic, TT-MES Temperate and tropical mesopelagic (divergent bathypelagic samples removed).

decay of viral particles by sunlight⁴² could also contribute to decoupling. However, at the depths sampled here (80 m; 200 m), sunlight-induced viral decay is likely to be lessened compared to surface samples (5 m) where previous turnover estimates have been calculated⁴¹.

A potential contributing factor for decoupling may be that the rate of active lytic viral replication in dominant members of the community such as SAR11 is low in the Sargasso Sea, as was previously observed in a coastal system¹⁶. This parallels a recent report that high abundances of free cyanophages coincided with low levels of *Prochlorococcus* infection in oligotrophic waters, proposed to result from a combination of loss of infectivity and low adsorption efficiency, alongside host resistance¹⁸. Success rates of SAR11 cultivation from dilution-to-extinction culturing experiments have been speculated to indicate that a large proportion of streamlined heterotrophs such as SAR11 and OM43 are possibly dormant, and therefore not contributing to viral turnover⁴³. In this scenario, viruses are released into the viral fraction but are unable to efficiently re-infect the cell-associated fraction, decoupling the two populations. High rates of lysogeny in taxa such as SAR11 would produce a similar outcome. Phylogenomic and genetic analysis of the abundant HTVC010P-type SAR11 pelagiphages has revealed that many encode an integrase gene (e.g., the isolate HTVC010P-type SAR11 prophage PNPI^{44,45}). However, the most abundant pelagiphages in the oceans do not encode known mechanisms of chromosomal integration^{14,15,45}. There was no evidence in environmental data collected on the cruise (C. Carlson, pers. comm.) to suggest entrainment of water from outside the gyre, nor do the high numbers of site-specific viral populations in the Sargasso Sea (Fig. 2) support possible decoupling by significant entrainment of viruses from outside the sample site concurrent to the sampling campaign.

Host prediction remains challenging despite MAGs

To determine viral taxonomic classification, we clustered the 2301 Sargasso Sea viral population representatives for genera-level classification based on shared-gene networks⁴⁶. RefSeq prokaryotic viral genomes (NCBI RefSeq v88 release) were included to assign family-level taxonomy to clusters. Out of 548 viral clusters, 186 contained

Sargasso Sea viral sequences, of which 177 lacked a representative sequence in the RefSeq database, consistent with previous studies of environmental viral communities^{31,47–50}. All phages assigned a taxonomy clustered with viruses within the order *Caudovirales* except one genome from family *Microviridae* with the order level *Petitvirales*. Few viral family-level taxonomic annotations were predicted: members of class Caudoviricetes (formerly known as 26 *Podoviridae*; 3 *Myoviridae*; 3 *Siphoviridae*); 1 *Microviridae*. Among the top 50 most abundant viral populations, 8% were classified, all of which were resolved as either *Podoviridae* or *Myoviridae*, highlighting the need for improved representation of abundant environmental viruses within reference databases⁵⁰. We next tried to assign putative hosts to Sargasso Sea viral populations by recovering metagenome-assembled genomes (MAGs) from cellular metagenomes. Assembled contigs were binned using sequence composition, relative abundance, and taxonomical classifications to group contigs into MAGs⁵¹. In total, we obtained 89 MAGs with $\geq 70\%$ completion and $\leq 10\%$ contamination that were used for host-prediction (Supplementary Dataset 3). Considering the intrinsic challenges of *in silico* host prediction^{52,53} a scoring matrix was developed to combine the results from prophage blast, tRNA scan, and WlsH to improve the accuracy of host assignments. Despite this, only six out of 2301 viral populations (0.26%) were successfully linked to three hosts which belonged to the phylum of Actinobacteriota ($n = 2$) and Chloroflexota ($n = 1$). Therefore, establishing virus-host linkages from metagenomes remains a major barrier for understanding environmental viral ecology.

Low abundance of known cyanophages and pelagiphages

During the summer months, cyanobacteria (*Prochlorococcus*) are the dominant phototrophs in the Sargasso Sea, with estimated relative abundances of up to 35% in the euphotic zone³⁴. Also highly abundant, SAR11 comprises 20–40% of the total bacterioplankton community in open ocean systems^{54,55}. Here, amplicon sequencing and analysis revealed high relative abundance of SAR11 at the sampling site over the course of the campaign. SAR11 16S rRNA Amplicon Sequence Variants (ASVs) contributed a maximum 52.7% and minimum 47% of the total amplicons from 80 m, and a maximum 45.6% and minimum 27.7% from 200 m (Supplementary Fig. 6). *Prochlorococcus* ASVs comprised a maximum 11.3% and minimum 4.1% of total amplicons from 80 m, and were not observed at 200 m (Supplementary Fig. 6). Previously, total viral abundance was shown to be negatively correlated with SAR11 abundance and positively correlated with *Prochlorococcus* over seasonal scales, leading to the hypothesis that viral communities were dominated by cyanophages with pelagiphages poorly represented¹¹. Composition analysis of the viral fraction here supports a dearth of pelagiphages throughout the sampling campaign. Recruitment of reads to *Pelagibacter* phage HTVC010P (a virus previously cited as the most abundant on Earth and isolated from the Sargasso Sea¹⁴, failed to meet the minimum genome coverage to be classified as present, even at a relaxed cut-off of 40% (Supplementary Fig. 7A, B). Moreover, *TerL* genes identified in the 100 most abundant viral populations did not cluster with known pelagiphage terminase genes within a phylogenetic tree (Supplementary Fig. 8). Thus, we found no evidence in either assemblies or short-read data of abundant viruses in the Sargasso Sea that are closely related to previously isolated pelagiphages. In comparison, the same read recruitment strategy revealed that known pelagiphages were well represented in global epipelagic viromes (GOV 2.0; TT-EPI), although less evident in mesopelagic samples (GOV 2.0; TT_MES) (Supplementary Fig. 7A, B). Conversely, no known pelagiphages were detected at $\geq 40\%$ minimum genome coverage in any cellular fraction metagenomes from the Sargasso Sea, either from this study or previously published cellular-fraction metagenomes⁵⁶ (Supplementary Fig. 7C, D).

Surprisingly, cyanophages were also rare in our Sargasso Sea samples. Among the top 100 most abundant viral populations, only four

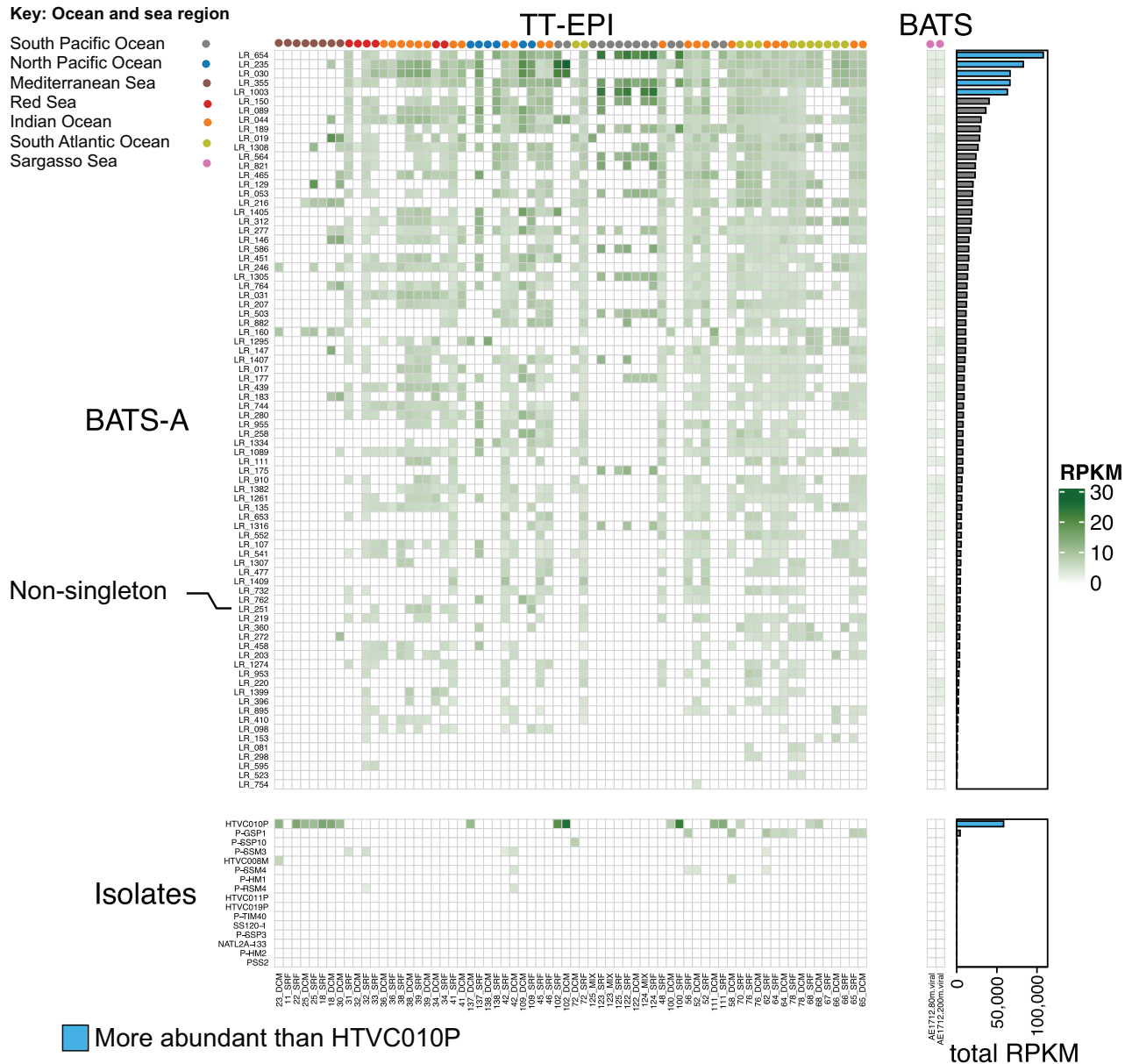


Fig. 4 | Global relative abundance of Sargasso Sea long-read viral contigs from this study (BATS-A) and isolates that infect SAR11 and *Prochlorococcus*. BATS-A viruses were important for discrimination between temperate and tropical epipelagic (TT-EPI) and temperate-tropical mesopelagic (TT-MES) viral populations in SIMPER analysis. 79 out of 80 BATS-A viruses were recovered from singleton populations (i.e., they were not captured by short-read assemblies); five of these

viral populations had greater global relative abundance (total RPKM) than Pelagiphage HTVC010P in the epipelagic (at middle and lower latitudes). The bootstrapped median ($n = 10,000$) number of Global Ocean Virome³⁰ (GOV2) samples in which a BATS-A viruses were observed was 56 (52.5–58 95% CI). Thus these viruses are common across oceanic regions, but were missed from existing short-read viral metagenomic datasets.

contigs were identified as potential cyanophages via presence of the *psbA* gene, with just two of these having completeness over 80%; these two populations were the 26th and 45th most abundant in the Sargasso Sea. In addition, when short read sequences from both the viral fraction and cellular fractions were recruited directly to a database of all published cyanophage and pelagiphage genomes, only two isolate genomes (viral fraction: *Synechococcus* phage Bellamy, MF351863.1; cellular fraction: *Prochlorococcus* phage P-SSM2, GU071092.1) were identified as present at a minimum genome coverage of 70% (Supplementary Fig. 7D). Although amplicon data showed that *Prochlorococcus* abundance did not approach the maximums recorded at this site (ref. 34, Supplementary Fig. 6), fewer cyanophages were observed than might be expected if active infection was prevalent during the sampling campaign. Low cyanophage abundance despite a comparatively higher

proportion of potential hosts contrasts with predictions by Parsons et al.¹¹. However, low cyanophage abundance may help to explain why Sargasso Sea viral populations from 80m and 200m were more similar to TT-MES samples (depth: 150–1000 m) in the GOV2 dataset (Fig. 3), where cyanophages are also rare, than TT-EPI samples (depth: 0–150 m), where cyanophages are abundant (Supplementary Fig. 7B). It is possible that known cyanophages or pelagiphages were not observed due to limitations in the sampling regime of this study (two depths, in one site over a consecutive four-day period) and that they may be detected at different times/depths/locations within the Sargasso Sea. Cyanophages were better represented in previously published cellular-fraction metagenomes from the Sargasso Sea³⁶ than in our samples (Supplementary Fig. 7C, D), suggesting the importance of sampling timing and duration within seasonal cycles.

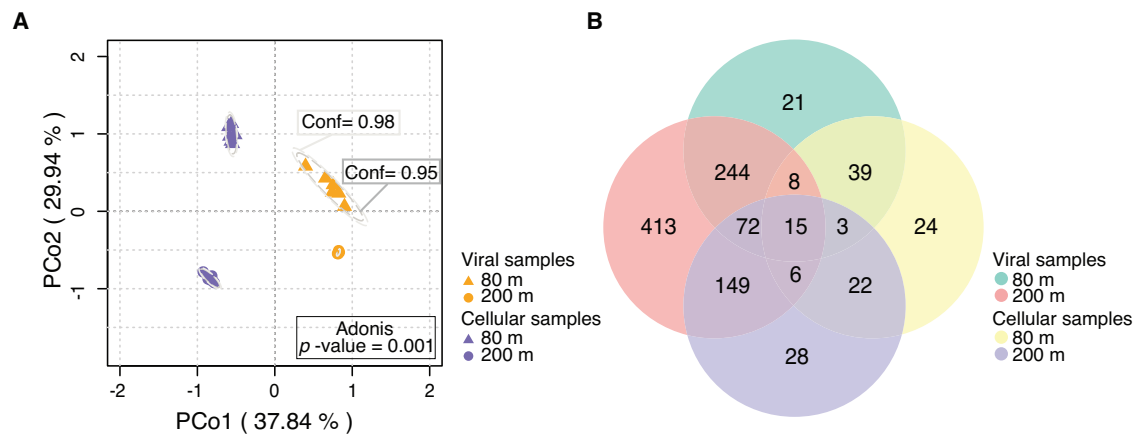


Fig. 5 | BATS viral populations clustered on the basis of sample fraction type (c = cellular; v = free particulate) and depth (80 m; 200 m). A Principal coordinates analysis (PCoA) of a Bray-Curtis dissimilarity matrix calculated from reads mapped to BATS viral populations ($n = 2301$) derived from short-read and VirION assemblies. Viral community structure was suggested by ellipses drawn at 95%

(inner) and 98% (outer) confidence intervals and analysis of variance (two-sided Adonis, p -value = 0.001). B Number of viral populations ($n = 1044$) found in 80m viral samples, 80m cellular samples, 200m viral samples, and 200m cellular samples, as identified from short-read only assemblies.

Rates of successful infection of *Prochlorococcus* in the oligotrophic surface waters of the North Pacific Subtropical Gyre (NPSG) were previously found to be low, despite a high abundance of cyanophage virions (~2.2% of the viral fraction)¹⁸. Therefore, it appears that in the NPSG, infection of a small proportion of the cyanobacterial community is sufficient to maintain abundant viral particles. The lack of representation of abundant pelagiphages and cyanophages observed in this study could suggest similarly low levels of active infection at this site in the Sargasso Sea over the duration of the sampling regime. Metabolic processes associated with pelagiphage infection were poorly represented in transcriptomic data from a coastal system collected over a 2-year period ($n = 8$), suggesting that chronic infection was more prevalent than lytic infection¹⁶. This, coupled with predicted dormancy rates of up to 85% in SAR11 cultured isolates⁴³ may further limit efficiency of pelagiphage infection in the Sargasso Sea. Yet in contrast to the NPSG study, we found little evidence of abundant cyanophages in the viral fraction from the Sargasso Sea at the time of sampling. An important difference between these regions is that phosphate concentrations in the Sargasso Sea are up to two orders of magnitude lower than those found in the NPSG⁵⁷. Synthesis of viral particles imposes a high phosphate requirement on infected cells and phosphate limitation restricts lytic infection in both cultures of cyanobacteria and natural marine communities (reviewed in ref. 58). We propose that during the sampling campaign, the increased P-limitation of the Sargasso Sea reduced viral production to a level where known pelagiphages and cyanophages (either isolates or closely related phage) were diluted to below levels of detection with metagenomic approaches. Here, clustering of the *TerL* genes identified in the 100 most abundant viral populations (Supplementary Fig. 8) implies that the most abundant Sargasso Sea viruses captured are most like those that infect copiotrophic hosts (e.g., *Pseudoalteromonas* and *Flavobacterium* phages). This could possibly suggest that boom-bust, particle associated interactions may be of greater importance than previously imagined in nutrient-limited water. However, as a result of the challenges of host-prediction, highlighted by our results here, this hypothesis is highly speculative at present, and more robust methods of host prediction are required for further investigation.

Sargasso Sea viral populations are microdiverse

Given that the Sargasso Sea bacterial community has more fine-scale, intraspecific diversity than variation at the species (or macrodiverse) level^{34,59}, we determined whether the same could be true of their

phages, and how levels of microdiversity might compare to those of global oceanic datasets. Here we report high microdiversity (i.e. intra-population diversity; π^{60}) values for Sargasso Sea viral populations across both depths (mean π : 3.411×10^{-4} ($2.473 \times 10^{-4} - 4.334 \times 10^{-4}$, 95% CI) (Supplementary Fig. 9), comparable to those recorded at similar latitudes in the Global Ocean Virome (GOV 2.0³⁰). The microdiversity of Sargasso Sea viral populations from 80 m and 200 m samples was not observed as significantly different (permutation and bootstrapping test: $p = 0.164$; Fig. 6A; Supplementary Fig. 10). This result does not align with those of Gregory et al.³⁰, who reported higher levels of microdiversity in mesopelagic temperate and tropical viromes than those from the epipelagic in the GOV 2.0 dataset. However, it is possible that the mesopelagic nature of the viral communities captured here work to reduce the signal of increasing microdiversity at depth observed in tropical and temperate viromes³⁰.

Additionally, we compared microdiversity of Sargasso Sea viral populations obtained using only short reads, and those captured with the inclusion of VirION reads, as previously we have observed that long reads assemble microdiverse viral genomes of Western English Channel viral assemblies better than short-reads due to the benefits of long-read assembly²³. The average microdiversity values for viruses derived from the two assembly types were 4.213×10^{-4} ($3.283 \times 10^{-4} - 5.326 \times 10^{-4}$, 95% CI) and 2.03×10^{-3} ($1.69 \times 10^{-3} - 2.36 \times 10^{-3}$, 95%CI), for short-reads and VirION reads, respectively (Fig. 6B), which represents an average increase of 389% (264.559 – 551.95%, 95% CI) in the π value calculated for viral genomes captured by VirION compared to those assembled from short-reads (permutation and bootstrapping significance test: $p < 0.001$) (Supplementary Fig. 11). Because π values are calculated from short reads which map to viral contigs, rather than the viral contigs themselves, this finding is not influenced by residual error in long-read derived contigs. This result confirms that VirION sequencing facilitates the capture of more viral microdiversity than is possible from short-read sequencing alone.

However, the question as to why oligotrophic regions such as the Sargasso Sea produce highly microdiverse viral assemblages remains open. The low infection rates observed in the NPSG¹⁸ suggests that phages here are not likely the main control on host abundance. Instead, such viruses may play an important role in the evolution of clonal, fine-scale diversity in the host population (as proposed decades ago¹⁷). Phages have been shown to increase host diversity on this micro scale in host-phage model systems⁶¹, and the high divergence in potential phage recognition sites observed in the HVRs of Pelagibacter

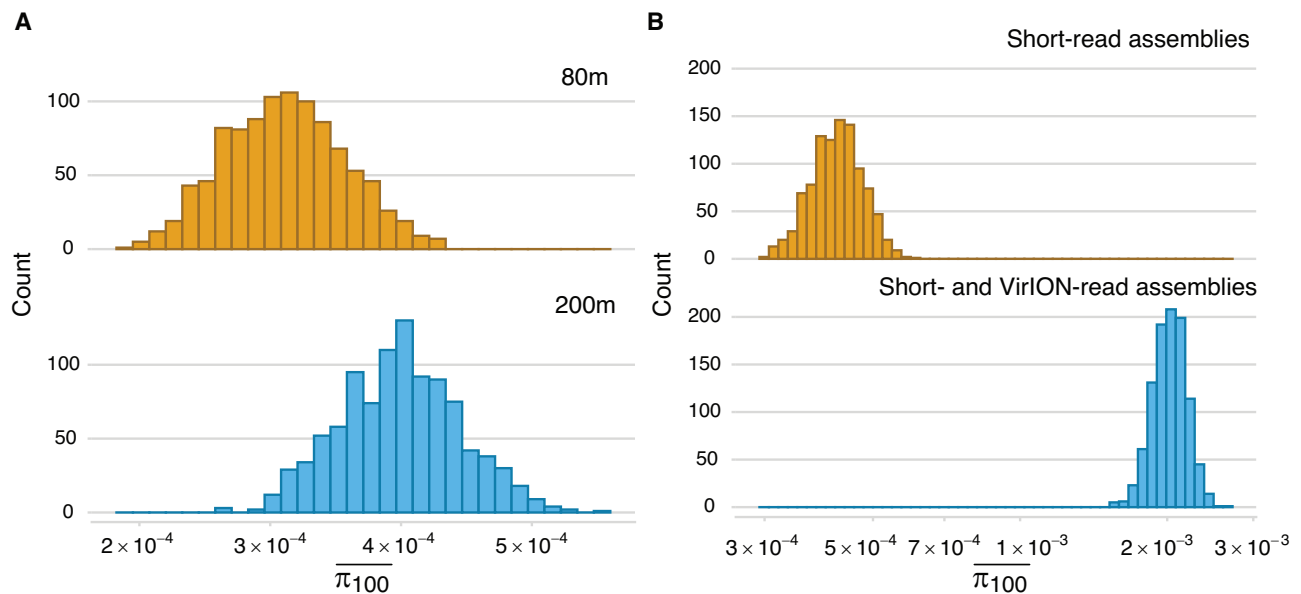


Fig. 6 | Microdiversity (intra-population diversity; mean π) of Sargasso Sea viral populations. (Calculated as Gregory et al.³⁰): **A** The microdiversity of Sargasso Sea viral populations from 200 m and 80 m samples (mean π : 3.119×10^{-4} (2.347×10^{-4} – 4.094×10^{-4} , 95% CI) and 3.967×10^{-4} (3.072×10^{-4} – 4.909×10^{-4} , 95% CI), respectively). On average, microdiversity was ~30% greater in viral genomes from 200 m than those from 80 m samples, but this was not a significant difference

(permutated significance test: $p = 0.164$; Supplementary Fig. 10B). **B** VirION (long-read assembly and hybrid long- and short-read assembly) facilitated the capture of viral genomes with high microdiversity: mean π was 388.668% (264.559–551.95%, 95% CI) greater for viral genomes captured by VirION ($n = 1465$) compared to those assembled from short-reads ($n = 2049$) (permutated significance: $p < 0.001$; Supplementary Fig. 11B).

isolates² supports the idea that phages drive host microdiversity in the Sargasso Sea. Because increased host microdiversity will expand phage niches, phage adaptation to emerging host ecotypes may increase phage microdiversity, as predicted by the Diversity Begets Diversity (DBD) theory of species interactions^{62,63}. DBD, which posits that existing diversity will promote the evolution of further diversity by niche construction (plus other types of interaction), appears to be most prevalent in low-diversity biomes⁶³.

Methods

Sample collection and DNA extraction

Metagenomic samples were collected aboard the *RV Atlantic Explorer* at the Bermuda Atlantic Time Series (BATS; <http://bats.bios.edu/>) station (31°40'N, 64°10'W) via rosette-mounted Niskin bottles during dusk (~19:00 local time) and dawn (~06:00 local time), from depths of 80 m and 200 m, over a period of four consecutive days from the 8th–11th July 2017. Host communities ($n = 12$) were obtained from 5 L of seawater per sample transferred immediately to clean polycarbonate bottles; the cellular fraction was recovered onto 0.22 μm pore Sterivex filters via positive pressure filtration. Each sample was stored in the dark at -20°C in 1 mL of SET buffer (0.75 M sucrose; 40 mM EDTA; 50 mM Tris-base). Within a fortnight of collection, DNA from the cellular fraction (which included host DNA, lysogenic viruses and any free viruses attached to cells) was extracted using phenol-chloroform^{64,65}, resuspended in 10 mM Tris-Cl buffer (pH 8.5), and stored at -20°C . Viral assemblages ($n = 12$) were obtained via sequential filtration of 20 L seawater per sample, followed by iron chloride flocculation⁶⁶, with modifications for prevention of DNA degradation and removal of PCR inhibitors²³. Briefly, peristaltic pumps and 142 mm rigs were used to remove the cellular fraction via sequential filtering through glass fibre (GF/D; pore size 2.7 μm) then polyethersulfone (pore size 0.22 μm) filters, before flocculation and precipitation of viruses via iron chloride. Iron-bound viral particle flocculate was recovered onto 1.0 μm polycarbonate filters (within 4 h of collection); filters were then stored in the dark at 4°C . Viruses were resuspended (within 5 months of collection) in ascorbate-EDTA buffer

(0.1 M EDTA, 0.2 M MgCl_2 , 0.2 M ascorbic acid, pH 6.1), concentrated using Amicon Ultra 100 kDa centrifugal filter units (Millipore UFC910024) and purified with DNase I (to remove un-encapsulated DNA). Viral DNA was extracted using the Wizard[®] DNA Clean-up System (Promega A7280). In contrast to previous samples derived from a coastal station (Western English Channel site L4²³; Sargasso Sea viral metagenomic DNA required further removal of PCR inhibitors prior to successful sequencing library preparation. This was accomplished using silica membrane spin columns (DNeasy PowerClean Pro Cleanup Kit; Qiagen 12997-50), before elution in 10 mM Tris-Cl buffer (pH 8.5), and storage at 4°C (colder storage temperatures were avoided to prevent freeze-thaw shearing of DNA for long-read sequencing library preparation).

Library preparation, amplification and sequencing

For short read sequencing, 1 ng of host community DNA and viral assemblage DNA was used to generate Nextera XT libraries (Illumina; manufacturer's protocol); After amplification (12 cycles) and assessment of library quantity (Qubit; ThermoFisher) and quality (Bioanalyzer; Agilent), library DNA was sequenced as 2×300 bp paired-end sequence reads, on a HiSeq 2500 (Illumina Inc.) in rapid mode, by the Exeter Sequencing Service (University of Exeter, UK). In addition, long-read sequences (mean average length ~4 kbp; Supplementary Table 2) were generated from viral metagenomic DNA via nanopore sequencing (Oxford Nanopore Technology: ONT) using the VirION pipeline²³, [dx.doi.org/10.17504/protocols.io.p8fdrtm](https://doi.org/10.17504/protocols.io.p8fdrtm). Briefly, ~100 ng of DNA per sample was sheared to ~8 kbp fragments to maximise PCR and sequencing efficiency and amplified using PCR-adapters for Linker Amplified Shotgun Library (LASL) generation. Samples from 200 m did not amplify sufficiently for preparation of long-read sequencing libraries, therefore downstream analysis focused on long-read assemblies and analysis of 80m samples. Three long read viral samples from 80m were prepared using the SQK-LSK109 kit and barcoded with native barcoding before being sequenced on a single MinION Mk 1B flowcell (FLO-MIN106; R9.4 SpotON; ONT)

Sequence processing and assembly

The following bioinformatic pipeline is summarised in Supplementary Fig. 1. Raw metagenomic short reads from both cellular and viral fractions were initially processed with cutadapt⁶⁷ to remove adaptors and PhiX reads (i.e., control library), then error-corrected and quality controlled with bmap (<https://jgi.doe.gov/data-and-tools/bbtools/>). Cleaned reads for each sample were assembled independently with metaSPAdes (v3.13.1⁶⁸ using k-mer sizes: 21, 33, 55 and 77. Long-read sequence data was basecalled with high accuracy using Guppy v3.3.0 and demultiplexed with Porechop (v0.4; <https://github.com/rrwick/Porechop>). Fewer than 42% of the reads were successfully assigned to a barcode, so following removal of adapters and barcodes with Porechop (including removal of reads with adapters in the middle), all reads from the three 80m samples were pooled together and filtered with NanoFilt⁶⁹ to remove those with a q-score < 10. Pooled, clean, high quality long reads were assembled (via Overlap Layout Consensus: OLC) with metaFlye⁷⁰ and Minipolish (including 10 rounds of Racon^{71,72}), followed by an additional polishing round with Medaka (<https://nanoporetech.github.io/medaka/>) and two rounds of short-read polishing with cleaned and pooled short reads from matched samples⁷³.

SAR11 and Prochlorococcus relative abundances

Metabarcoding 16S rRNA amplicon sequencing and analysis was performed as follows: Four litres of seawater were filtered onto 0.2 µm Sterivex filters. DNA was extracted using a phenol-chloroform protocol⁷⁴ and amplified with V1-V2 primers 27F (5'-AGAGTTTGGATCNTGGCTCAG-3') and 338RPL (5'-GCWGCCWCCCGTAGGWTG-3'). Libraries were sequenced using 2 × 250 Pair-End on the MiSeq platform (Reagent Kit v2). The 16S rRNA sequences were trimmed and dereplicated, chimera checked, and ASVs generated using the DADA2 R⁷⁵ package, version 1.27⁶. Taxonomy was assigned with the Silva database version 123⁷⁷. Amplicon datasets collected at 80 m and 200 m depth from the BIOS-SCOPE Cruise AE1712 core casts were extracted with Phyloseq⁷⁸ and relative contributions were plotted using the ggplot package⁷⁹ in R⁷⁵ (above pipeline as described by ref. 80). Amplicon raw data SRAs used in this study are available at NCBI Bioproject PRJNA769790.

Viral sequence recovery and dereplication

To identify Sargasso Sea viruses in our assemblies, contiguous sequences from long and short read assemblies that were circular or ≥ 10 kb were processed with VirSorter (v1.0.5²⁹) after augmenting its database with the Xfam database of viral HMM profiles from Guo et al.⁸¹; contigs resolved into categories 1, 2, 4, and 5 were classed as putatively viral and passed into downstream analyses²⁴. One viral sample (collected July 8th from 200 m) was excluded from further investigation due to a lack of viral sequence detection, the result of very low sequencing depth (29.6 Mbp, 100 times smaller than that of other samples). Viral contigs were dereplicated into viral populations by clustering those that shared ≥ 95% nucleotide identity across ≥ 85% of the contig length, using ClusterGenomes (<https://github.com/simroux/ClusterGenomes>)²⁴. The longest contig within a cluster was selected as the cluster (population) representative. Where populations were represented by contigs derived from long reads, the presence of fragmentation in short-read assemblies was investigated by mapping: I) population members derived from short reads; II) all short-read contigs >1 kb (Minimap2⁸²). The results of mapping were parsed with a custom-written Python script to identify regions of the representatives that were not covered by short-read assemblies ('calc_breakages.py'; available from <https://github.com/BIOS-SCOPE/AE1712-viromes>). Using the R⁷⁵ packages ggplot⁷⁹, tidyverse⁸³, cowplot⁸⁴, colorspace⁸⁵ and ggrepel⁸⁶, fragmentation (number of breakages) and recovery (percentage aligned) of the representatives were plotted against their relative abundances (RPKM values in short-read data calculated with

CoverM (v0.2.0) (<https://github.com/wwood/CoverM>)). Cluster representatives were used in a second round of clustering with a combined dataset of Global Ocean Viromes 2.0 dataset (GOV 2.0³⁰; accession numbers ENA: PRJEB402; PRJEB9742; NCBI: PRJNA366219) to identify viruses that belonged to known viral populations in Global Ocean datasets and those that were novel to the Sargasso Sea.

Metagenome assembled genomes

To increase likelihood of host-prediction of phage contigs, the genomes of cellular Sargasso Sea microbes were assembled as follows: Short reads were mapped to the contigs from the cellular fraction using CoverM (v0.2.0) (<https://github.com/wwood/CoverM>). Reads were retained if they had > 95% identity and > 75% read coverage with trimmed_mean used to remove the top 5% and bottom 5% depths. The bam files from read-mapping and the contigs from cellular samples were used as inputs into a custom script for binning. First, UniteM (v0.0.18) (UniteM; unpublished <https://github.com/dparks1134/UniteM>) was used to create a set of initial bins using the unitem bin command, utilizing the following binning options: gm2, bs, mb2, max40, max107, mb_very-sensitive, mb_sensitive, mb_specific, mb_very-specific. From this set of bins, UniteM profile and UniteM consensus commands were used to produce an ensemble bin set. Concordantly, DAS Tool (v1.1.1)⁸⁷ and the MetaWRAP (v1.0.6)⁸⁸ bin_refinement module were utilized on the bins produced from the unitem bin command to produce ensemble bins. However, the MetaWRAP (v1.0.6) bin_refinement module only accepts three candidate bin sets, so MetaBAT2⁸⁹, GroopM2⁹⁰, and MaxBin2⁹¹ outputs from UniteM were used as input into the MetaWRAP module. After all ensemble binning techniques were complete (MetaWRAP, DAS Tool, UniteM), the product ensemble bins were used as input for UniteM, MetaWRAP, and DAS Tool for a second iteration to produce an optimal bin set. Following the second iteration of ensemble binning, the output bins of each tool individually was evaluated for completeness and contamination using the CheckM (v1.0.12)⁹² lineage workflow. Once the completeness and contamination statistics for the bin sets of the second iteration of ensemble binning tools were obtained, the bins greater than or equal to 70% completion and less than or equal to 10% contamination were used to calculate a quality score. Similar to the methods used in UniteM and CheckM, a score was calculated as the following: score = completeness - (2 × contamination). Each ensemble binning tool was scored, and then the tool with the highest quality score was used as the bin set for that particular sample. Following scoring, RefineM (v0.0.24)⁹³ outliers was used to remove any potential outliers associated with GC content or tetranucleotide signatures with the following parameters: -gc_perc 95 -td_perc 95. The taxonomical classification of the resulting 89 MAGs was undertaken with GTDB-tk18⁹⁴; genomes were classified via placement in a GTDB reference tree.

Local and global relative abundance calculations

Next, we investigated which viruses were most abundant in our samples, and how Sargasso Sea viruses are distributed across the Global Ocean. For local abundance calculations, Illumina reads that passed quality controls were competitively recruited to the Sargasso Sea dataset of viral population representatives (derived from both VirIION and short-read assemblies) using Bowtie2⁹⁵, in non-deterministic, sensitive mode; the resulting bam files were parsed in BamM (<https://github.com/ECogenomics/BamM>) to retain reads that mapped at ≥ 90% read length at ≥ 95% identity²⁴. The abundance of viral populations within each sample were calculated using mean contig coverage (excluding <5th and >95th percentile - tpm³⁰ using BamM coverage. Population representatives with < 70% coverage³⁰ within a sample were assigned an abundance of 0 to minimise false positive detection of populations within a sample²⁴. Coverage values of viral populations were then normalized by total number of reads per metagenome as a proxy for relative abundance. Total relative abundance of long-read derived and short-read derived viral populations, and differences

between the lengths of those viral populations, was tested and plotted with the R⁷⁵ packages tidyverse⁸³, cowplot⁸⁴, and scales⁸⁶. Local diversity of the Sargasso Sea viral populations from cellular fraction and viral fraction samples was assessed using the packages vegan⁹⁷ (<https://CRAN.R-project.org/package=vegan>) and pracma (<https://cran.r-project.org/web/packages/pracma/index.html>) in R⁷⁵. Community structure of Sargasso Sea viruses was evaluated with Principal Coordinates Analysis (PCoA) based on Bray-Curtis dissimilarity (function vegdist) from cube-root transformed abundance of representative viral contigs. Clusters were tested for statistical significance using standard deviation and F-tests (function Adonis; 999 permutations); ellipses were visualised (function ordiellipse) at 95% (inner) and 98% (outer) confidence intervals. Presence-absence of viral populations in cellular fraction and viral fraction samples was visualised using Adobe Illustrator⁹⁸.

To investigate the global distribution of the recovered Sargasso Sea viral populations, short reads from this study and from the GOV 2.0 dataset³⁰ were mapped back to the Sargasso Sea and GOV2 representative viral population contigs. The GOV 2.0 dataset contains 145 samples from five distinct global ecological zones, including the Arctic, Antarctic, bathypelagic, temperate and tropical epipelagic, and mesopelagic. Datasets were subsampled to 5 million reads prior to read mapping to prevent sequencing depth influencing the likelihood of contigs meeting the minimum genome coverage cut-off value ($\geq 70\%$) and thus possible inflation of the number of rare viruses detected as present in larger datasets compared to smaller datasets. Subsampled reads were recruited against a dereplicated set of Sargasso Sea and GOV2 viral population contigs (using the cut-offs and read recruitment strategy detailed above). Estimated presence/absence values of Sargasso Sea viruses were then calculated singly for each GOV2 site, and for the dataset as a whole. Presence/absence of Sargasso Sea viral populations in the GOV2 dataset were plotted using the ggplot package⁷⁹ in R⁷⁵, as were normalised (between sample sites) levels of nitrogen (NO₃ = Nitrate+Nitrite-1 (umol/kg) and Phosphate (PO₄ = Phosphate-1 (umol/kg)). Global distribution of Sargasso Sea samples was visualised with the R⁷⁵ packages Simple Features⁹⁹ and rnaturalearth (<https://github.com/ropensci/rnaturalearth>), and the vector graphics editor Inkscape¹⁰⁰.

Viral classification, microbial survey and host prediction

Having examined the abundance of Sargasso Sea viruses, we next investigated which viruses were present, how the bacterial community was structured, and whether we could predict the hosts of our viral genomes. To classify viruses, open reading frames (ORFs) in viral population representatives were identified with Prodigal (v2.6.1)¹⁰¹ in metagenomic mode with default settings. Population representatives were clustered into ICTV-recognized viral genera using vConTACT2¹⁰² alongside RefSeq prokaryotic viral genomes (release 88) for reference to known isolates.

Linkage of viral populations to putative hosts was attempted as follows: Putative hosts were assigned to viral populations through prophage blast, tRNAscan-SE (v1.23)¹², and WisH (v1.0)¹³ using a scoring approach similar to what has been previously reported for human gut viromes¹⁴. In prophage blast, a nucleotide blast database was built by using Sargasso Sea MAGs. Viral representative contigs were used as input to BLAST against this database. Scores ranging from 1 to 4 were assigned based on percent identity and coverage (4: 98% ID and 90% cov, 3: 90% ID and 75% cov, 2: 90% ID and 50% cov, 1: 90% ID and 30% cov). General tRNA models were predicted for viral contigs in tRNAscan-SE (v1.23). Secondary structures of MAGs were searched using bacterial tRNA model. Scores were assigned to hits according to percent identity (3: 100%, 2: 95%, 1: 90%). Host models were built in WisH (v1.0). Null models were predicted by using 283 decoy RefSeq viral sequences that infect non-marine isolates belonging to the genera Staphylococcus, Streptococcus, Lactobacillus, Propionibacterium,

Mannheimia, and Paenibacillus since none of these genera should encompass ocean MAGs. The virus-host linkages were predicted by providing target viral contigs, host MAG models, and the matrix of null model parameters. Scores were given based on reported p-values (p -value $\leq 10^{-10}$: 2.5, p -value $\leq 10^{-5}$: 2). Collectively, virus-host linkages that had scores ≥ 3 were considered as putative hosts. As just 0.26% of viral contigs were successfully linked to hosts, alternative methods of host prediction were then attempted.

Recalling previous evidence that at Sargasso Sea, *Prochlorococcus* viruses may be more important to total viral abundance than SAR11 viruses (*sensu*¹¹), we evaluated evidence of these phages in Sargasso Sea viromes using two approaches: First, we determined if assembled contigs from Sargasso viromes could be associated with SAR11 or *Prochlorococcus* hosts. The top 100 most abundant viral populations were screened for marker genes associated with cyanophages and pelagiphages using DRAM-v¹⁰³. Specifically, contigs containing photosynthetic *psbA* gene (prevalent in cyanophages)¹⁰⁴ were extracted as putative cyanophages for manual curation and assessed for completeness with CheckV v0.3.0¹⁰⁵. Cultured pelagiphages lack appropriate signature auxiliary metabolic genes which could putatively be used to identify pelagiphages from metagenomic data. However, terminase (*TerL*) genes are commonly used to construct pelagiphage phylogenies^{13,106} which correlate to clustering of pelagiphage genomes from shared-genenetworks (e.g. vConTACT2)¹⁰⁶. To identify pelagiphages, DRAM-v annotations that specified terminase (*TerL*) genes were identified from published pelagiphage genomes, non-pelagiphages (from NCBI refseq, search term: 'marine terminase in viruses'), *Pelagibacterales* (identified via BLAST hits against terminase from known pelagiphages) and Sargasso Sea viral populations. *TerL* genes were aligned using E-INS-i strategy for 1000 iterations in MAFFT v7.017¹⁰⁷. Aligned sequences were trimmed with Trimal v1.4.rev15¹⁰⁸ with sites containing more than 50% gaps removed from the alignment. Alignments were manually checked for overhangs in Geneious v10.2.6¹⁰⁹. After determining an appropriate substitution model using Model Finder¹¹⁰, a phylogenetic tree was constructed using IQ-tree¹¹¹ with rapid bootstrap support generated from 1000 iterations. The phylogeny was visualized in iTOL v5¹¹², and each clade was subsampled to improve clarity whilst retaining diversity within the tree. Sargasso Sea populations which clustered more closely to non-pelagiphages than pelagiphages were considered to demonstrate dissimilarity to known viruses of SAR11.

Next, we assessed whether genomes similar to those of previously isolated viruses of cyanobacteria and SAR11 were represented in Sargasso Sea viromes through mapping of short reads. High quality short reads were mapped against a dereplicated set of all published cyanophage and pelagiphage isolate genomes (accession numbers: Supplementary Dataset 4). Viral population representatives were generated using ClusterGenomes²⁴ as above (cut-offs $\geq 95\%$ nucleotide identity across $\geq 85\%$ genome length). *Escherichia* phage T4 was added as a negative control. Reads were recruited using Bowtie2⁹⁵, in non-deterministic, sensitive mode, and the resulting bam files were parsed in CoverM (<https://github.com/wood/CoverM>) to retain reads that mapped at $\geq 90\%$ read length at $\geq 95\%$ identity²⁴. RPKM was calculated as a proxy for relative abundance. To evaluate whether detection of phage isolates was sensitive to minimum genome coverage cutoffs, we conducted this analysis using minimum coverage thresholds of 40% and 70%. Lastly, to compare the representation of previously isolated viruses of cyanobacteria and SAR11 in Sargasso Sea viromes to previously published Sargasso Sea samples and global viromes, we repeated this analysis using previously published metagenomes from the Sargasso Sea⁵⁶ (cellular-fraction only; no quantitative viromes of dsDNA viruses from the Sargasso Sea has been previously published), and the GOV 2.0 dataset. Data manipulation and plotting was conducted in R⁷⁵, using packages tidyverse⁸³, cowplot⁸⁴, and colorspace⁸⁵.

Inter-community diversity calculations

Evaluation of the inter-community diversity of the Sargasso Sea viral populations in relation to global viral populations was conducted as follows: Filtered and sequencing-depth normalized read mappings of the complete GOV2 dataset and Illumina reads from BATS against a combined, dereplicated database of Sargasso Sea (this study) and GOV 2.0³⁰ viral populations ($n=152,979$) for abundance calculations (above) were processed to evaluate viral inter-community diversity using the *vegan* package⁹⁷ (<https://CRAN.R-project.org/package=vegan>) in R⁷⁵. Viral β -diversity and community structure was evaluated with PCoA based on Bray-Curtis dissimilarity (function *vegdist*) from cube-root transformed abundance of representative viral contigs. Statistical support for clusters was evaluated using standard deviation and F-tests (function *Adonis*; 999 permutations) and visualised with ellipses drawn at 95% (inner) and 98% (outer) confidence intervals (function *ordiellipse*). The *PairwiseAdonis* command from package *pairwiseAdonis*¹¹³ in R⁷⁵ was used to generate pairwise comparisons between the groups: centroids were calculated from the mean of the samples' weighted averages within each group along the unconstrained ordination axes, and Euclidean distances were estimated for each pair of centroids. A Similarity Percentages (SIMPER) analysis (function *adonis*) was conducted to identify the key Sargasso Sea viral populations driving the community structure depicted via PCoA. To ascertain if the most important contributors to β -diversity were comprised of globally abundant or Sargasso-Sea specific viruses, a bootstrap ($n=10000$) test was conducted of the median number of Global Ocean Virome (GOV2) samples in which each Sargasso Sea virus appeared (code for bootstrapping available from: [https://raw.githubusercontent.com/btemperton/tempertonlab_utils/master/R/StatsUtilities.R]¹¹⁴). Global relative abundance of Sargasso Sea long-read viral contigs and isolates that infect SAR11 and *Prochlorococcus* were plotted using R⁷⁵ packages *tidyverse*⁸³, *cowplot*⁸⁴, *scales*⁹⁶, and *ComplexHeatmap*¹¹⁵, with editing in *Inkscape*¹⁰⁰ for addition of Ocean and sea regions.

Microdiversity calculations

We next evaluated microdiversity in Sargasso Sea virome short- and long-read data to investigate: 1) whether the nutrient-limited waters of the Sargasso Sea were enriched in microdiverse viral populations typically associated with hosts that favour such conditions; 2) whether microdiversity was significantly different between viral populations from 80 m and 200 m; 3) whether previous findings that long-read viromes capture greater microdiversity in the viral fraction in a coastal region were similarly observed in viromes from nutrient-depleted environments^{23,26}. To compare average microdiversity (nucleotide diversity: π ⁶⁰; between in Sargasso Sea viruses and those captured in the Global Ocean Virome (GOV2), we used the same approach as Gregory et al.³⁰. Long-reads (not available in GOV2 datasets) were excluded from this analysis to avoid potential artefacts associated with sequencing technology. Briefly, all short-read Sargasso Sea viromes were randomly subsampled without replacement to 1M reads using *bbmap* (<https://sourceforge.net/projects/bbmap/>). The subsampled reads were assembled, viruses identified, and relative abundances calculated (all methods as above) to generate BAM files. BAM files extracted from *Bowtie2* with default parameters were used as inputs to *metapop*¹¹⁶ to call single nucleotide variants (SNVs). Viral populations were only included if $\geq 70\%$ of representative contigs were covered with an average depth of $\geq 10X$. SNVs with a quality call of > 30 (QUAL score; phred-scaled) were retained, and only those with alternative alleles with a frequency $> 1\%$ and supported by ≥ 4 reads were regarded as SNV loci. To minimize sequencing errors and address coverage variations, coverage was randomly subsampled to 10X coverage per locus across the genome. To calculate average microdiversity in Sargasso Sea viruses for comparison to Global Ocean Virome (GOV2) and between depths, mean π from 80m and 200m samples were calculated

with 1000 bootstraps of 100 randomly subsampled π values from short-read Sargasso Sea viromes (with replacement). Differences in mean π between depths in Sargasso Sea samples were compared to a null model in which π values from both depths belong to one population (by randomly shuffling of labels and splitting into two datasets of equal size to initial datasets). To investigate whether microdiversity in VirION-derived viral genomes was significantly higher than those in short-read only viral genomes, preliminary work included the generation of additional BAM files by mapping short reads to short-read and VirION assemblies (together; method as above), and the calculation of 100 permuted π values from Sargasso Sea viruses assembled using each approach (i.e., short reads only/VirION pipeline) to generate distributions within those permutations. The permuted percentage increase between the mean π values from each approach was then tested for significance via permutation and bootstrapping test (1000 iterations), as above. The R⁷⁵ packages used for data manipulation, permutation and bootstrapping tests and plotting results were *Tidyverse*⁸³, *Cowplot*⁸⁴, *Colorblindr*(<https://rdocumentation.org/packages/colorblindr/versions/0.1.0>), *Colorspace*⁸⁵, and *Scales*⁹⁶.

Investigation of virus hypervariable regions

To investigate the presence and content of hypervariable regions, high quality short reads were mapped to the top 50 most abundant Sargasso Sea viral population representatives (derived from both VirION and short-read assemblies) using *Bowtie2*⁹⁵. The resulting bam files were filtered to retain quality mappings (at 95% identity and 70% coverage) using *CoverM* (<https://github.com/wwood/CoverM>). The per-nucleotide coverage of the top 50 most abundant viruses was generated from the filtered bam files using *bedtools2* (<https://github.com/arq5x/bedtools2>) and used as input to find HVRs, which were defined as regions that possessed: less than 20% of the whole contig median coverage; at least 600 bp; zones of zero coverage¹¹⁷. Functions encoded within candidate HVR regions were investigated using a tBLASTx search against the NCBI NR database.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequencing data and assemblies generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) database under the BioProject accession code [PRJNA767318](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA767318). The amplicon raw SRA data used in this study are available in the NCBI database under the BioProject accession code [PRJNA769790](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA769790). The following data generated in this study are provided in the Supplementary Information: PERMANOVA analysis results (Supplementary Dataset 1); 'BATS-B' (from SIMPER analysis) contig abundance (Supplementary Dataset 2); Metagenome Assembled Genomes (MAGs) (Supplementary Dataset 3); Accession numbers of published cyanophage, pelagiphage and T4 genomes used in the production of Supplementary Fig. 7 (Supplementary Dataset 4). The following data generated in this study are provided in the Source Data file: Alignments of Sargasso Sea long-read derived viral population representatives against short-read derived population members; used as input for the script *calc_breakages.py*, towards production of Supplementary Figs. 2A and B (alignment_short_members_tO_long_reps.txt); Areas of Sargasso Sea long-read derived viral population representatives not aligned to short-read population members; output of *calc_breakages.py*, towards production of Supplementary Fig. 2A, B (LRR_breakages_2023_02_23.csv); Relative abundance (RPKM) of Sargasso Sea long-read derived viral population representatives in short-read data; towards production of Supplementary Fig. 2A, B (long_read_cluster_rep_rel_abundance_covminzero.txt); Alignments of Sargasso Sea long-read derived viral population representatives against short-read contigs $< 1\text{kb}$ in length;

used as input for the script `calc_breakages.py`, towards production of Supplementary Fig. 2C (`alignment_shortread_contigs_gr1kb_to_all_LRR-sorted-by-target-name.txt`); Areas of Sargasso Sea long-read derived viral population representatives not aligned to short-read contigs <1kb; Output of `calc_breakages.py`; towards production of Supplementary Fig. 2C (`breakages_all_LRR_contigs_gr_1kb.csv`); Relative abundance (RPKM) of Sargasso Sea long-read derived viral population representatives in short-read data; towards production of Supplementary Fig. 2C (`All_long_read_cluster_rep_RPKM.txt`); Abundance tables generated from mapping Sargasso Sea reads and GOV2 reads to Sargasso Sea population representative contigs, plus GOV2 metadata; used in production of Figs. 3 and 5A. (folder `figure5a_figure3`); Microdiversity values for Sargasso Sea contigs used to produce Supplementary Figs. 9, 10, and 11 (`sl_contig_microdiversity.tsv`); Viral population representative contig names, lengths and type; used for production of Fig. 1 (`length_2301.txt`); Rank and abundance values for short-read viral population representatives; used for production of Figs. 1, 5 (`rank_abundance_2301.csv`); Metadata (including labels) for GOV2 and Sargasso Sea samples; used for production of Fig. 3 (`BATS_GOV2_0_env.csv`); Short-read coverage of GOV2 and Sargasso Sea viral populations; used for production of Fig. 3 (`GOV2_0_BATS_coverage`); Metadata (including labels) for Sargasso Sea samples; used for production of Fig. 5 and Supplementary Figs. 3, 4 and 5 (`BATS_env.csv`); Short-read coverage of Sargasso Sea viral populations; used for production of Fig. 5A (`BATS_short_reads_2031_coverage_rm2v.csv`); Short-read coverage of Sargasso Sea viral populations; used for production of Supplementary Fig. 3A (`GOV2_0_BATS_coverage_S3A.csv`); Short-read coverage of Sargasso Sea viral populations; used for production of Supplementary Fig. 3B (`GOV2_0_BATS_coverage_S3B.csv`), Short-read coverage of Sargasso Sea viral populations from short-read sequencing; used for production of Supplementary Fig. 4 (`BATS_short_reads_1044_coverage.csv`) Source data are provided with this paper.

Code availability

Code for analyses and associated data used in the current study are available at: <https://github.com/BIOS-SCOPE/AE1712-viromes>, <https://doi.org/10.5281/zenodo.10940125>¹¹⁸.

References

- Suttle, C. A. Marine viruses—major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
- Rodriguez-Valera, F. et al. Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).
- Lindell, D. et al. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**, 83–86 (2007).
- Lindell, D. et al. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. USA* **101**, 11013–11018 (2004).
- Scanlan, P. D. et al. Coevolution with bacteriophages drives genome-wide host evolution and constrains the acquisition of abiotic-beneficial mutations. *Mol. Biol. Evol.* **32**, 1425–1435 (2015).
- Forterre, P. The virocell concept and environmental microbiology. *ISME J.* **7**, 233–236 (2012).
- Howard-Varona, C. et al. Phage-specific metabolic reprogramming of virocells. *ISME J.* **14**, 881–895 (2020).
- Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
- Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J. & Temperton, B. Host-hijacking and planktonic piracy: How phages command the microbial high seas. *Viral. J.* **16**, 1–13 (2019).
- Guidi, L. et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
- Parsons, R. J., Breitbart, M., Lomas, M. W. & Carlson, C. A. Ocean time-series reveals recurring seasonal patterns of viroplankton dynamics in the northwestern Sargasso Sea. *ISME J.* **6**, 273–284 (2012).
- Kang, I., Oh, H.-M., Kang, D. & Cho, J.-C. Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proc. Natl. Acad. Sci. USA* **110**, 12343–12348 (2013).
- Zhang, Z. et al. Culturing novel and abundant pelagiphages in the ocean. *Environ. Microbiol.* **00**, 1–17 (2020).
- Zhao, Y. et al. Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360 (2013).
- Martinez-Hernandez, F. et al. Single-cell genomics uncover *Pelagibacter* as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. *ISME J.* **13**, 232–236 (2019).
- Alonso-Sáez, L., Morán, X. A. G. & Clokie, M. R. Low activity of lytic pelagiphages in coastal marine waters. *ISME J.* **12**, 2100–2102 (2018).
- Waterbury, J. B. & Valois, F. W. Resistance to co-occurring phages enables marine *Synechococcus* communities to coexist with cyanophages abundant in seawater. *Appl. Environ. Microbiol.* **59**, 3393–3399 (1993).
- Mruwat, N. et al. A single-cell colony method reveals low levels of infected *Prochlorococcus* in oligotrophic waters despite high cyanophage abundances. *ISME J.* **15**, 41–54 (2021).
- Kelly, L., Ding, H., Huang, K. H., Osburne, M. S. & Chisholm, S. W. Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *ISME J.* **7**, 1827–1841 (2013).
- Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F. & Chisholm, S. W. Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biol.* **3**, 0790–0806 (2005).
- Sullivan, M. B. et al. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**, 3035–3056 (2010).
- Buchholz, H. H. et al. A Novel and Ubiquitous Marine Methylophage Provides Insights into Viral-Host Coevolution and Possible Host-Range Expansion in Streamlined Marine Heterotrophic Bacteria. *Appl. Environ. Microbiol.* **88**, e0025522 (2022).
- Warwick-Dugdale, J. et al. Long-read viral metagenomics enables capture of abundant and microdiverse viral populations and their niche-defining genomic islands. *Peer J.* **7**, e6800 (2019).
- Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *Peer J.* **5**, e3817 (2017).
- Zablocki, O. et al. VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. *Peer J.* **9**, e11088 (2021).
- Martinez-Hernandez, F. et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).
- Jacquet, S., Partensky, F., Lennon, J. F. & Vaulot, D. Diel patterns of growth and division in marine picoplankton in culture. *J. Phycol.* **37**, 357–369 (2001).
- Ottesen, E. A. et al. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science (1979)* **345**, 207–212 (2014).
- Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
- Gregory, A. C. et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109–1123.e14 (2019).
- Angly, F. E. et al. The marine viromes of four oceanic regions. *PLoS Biol.* **4**, 2121–2131 (2006).

32. Morris, R. M. et al. Temporal and spatial response of bacterioplankton lineages to annual convective overturn at the Bermuda Atlantic Time-series Study site. *Limnol. Oceanogr.* **50**, 1687–1696 (2005).
33. Carlson, C. A. et al. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J.* **3**, 283–295 (2009).
34. Treusch, A. H. et al. Seasonality and vertical structure of microbial communities in an ocean gyre. *ISME J.* **3**, 1148–1163 (2009).
35. Vergin, K. L. et al. High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J.* **7**, 1322–1332 (2013).
36. Giovannoni, S. J. & Vergin, K. L. Seasonality in ocean microbial communities. *Science* (1979) **335**, 671–676 (2012).
37. Trubl, G. et al. Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems* **3**, 1–21 (2018).
38. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
39. Santos-Medellin, C. et al. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J.* **15**, 1956–1970 (2021).
40. Schroeder, P. J. & Jenkins, D. G. How robust are popular beta diversity indices to sampling error. *Ecosphere* **9**, (2018).
41. Noble, R. T. & Fuhrman, J. A. Rapid virus production and removal as measured with fluorescently labeled viruses as tracers. *Appl. Environ. Microbiol.* **66**, 3790–3797 (2000).
42. Suttle, C. A. & Chen, F. Mechanisms and rates of decay of marine viruses in seawater. *Appl. Environ. Microbiol.* **58**, 3721–3729 (1992).
43. Henson, M. W. et al. Expanding the diversity of bacterioplankton isolates and modeling isolation efficacy with large scale dilution-to-extinction cultivation. *Appl. Environ. Microbiol.* **86**, e00943–20 (2020).
44. Morris, R. M., Cain, K. R., Hvorecny, K. L. & Kollman, J. M. Lysogenic host–virus interactions in SAR11 marine bacteria. *Nat. Microbiol.* **5**, 1011–1015 (2020).
45. Du, S. et al. Genomic diversity, life strategies and ecology of marine htvC010p-type pelagiphages. *Micro. Genom.* **7**, 000596 (2021).
46. Bolduc, B. et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243 (2017).
47. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus – host interactions resolved from publicly available microbial genomes. *Elife* **4**, 1–20 (2015).
48. Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res* **239**, 136–142 (2017).
49. Hurwitz, B. L. & Sullivan, M. B. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLoS One* **8**, e57355 (2013).
50. Paez-Espino, D. et al. Uncovering Earth’s virome. *Nature* **536**, 425–430 (2016).
51. Gregory, A. C. et al. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724–740.e8 (2020).
52. Moon, K. & Cho, J. Metaviromics coupled with phage-host identification to open the viral ‘black box’. *J. Microbiol.* **59**, 311–323 (2021).
53. Khot, V., Strous, M. & Hawley, A. K. Computational approaches in viral ecology. *Comput Struct. Biotechnol. J.* **18**, 1605–1612 (2020).
54. Morris, R. M. et al. SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806–810 (2002).
55. Rappé, M. S. & Giovannoni, S. J. The uncultured microbial majority. *Annu Rev. Microbiol.* **57**, 369–394 (2003).
56. Biller, S. J. et al. Marine microbial metagenomes sampled across space and time. *Sci. Data* **5**, 180176 (2018).
57. Wu, J., Sunda, W., Boyle, E. A. & Karl, D. M. Phosphate Depletion in the Western North Atlantic. *Ocean. Sci.* (1979) **289**, 759–762 (2020).
58. Wilson, W. H. & Mann, N. H. Lysogenic and lytic viral production in marine microbial communities. *Aquat. Microb. Ecol.* **13**, 95–100 (1997).
59. Venter, J. C. et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* (1979) **304**, 66–74 (2004).
60. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273 (1979).
61. Brockhurst, M. A., Buckling, A. & Rainey, P. B. The effect of a bacteriophage on diversification of the opportunistic bacterial pathogen, *Pseudomonas aeruginosa*. *Proc. R. Soc. B: Biol. Sci.* **272**, 1385–1391 (2005).
62. Calcagno, V., Jarne, P., Loreau, M., Mouquet, N. & David, P. Diversity spurs diversification in ecological communities. *Nat. Commun.* **8**, 1–9 (2017).
63. Madi, N., Vos, M., Murall, C. L., Legendre, P. & Shapiro, B. J. Does diversity beget diversity in microbiomes? *Elife* **9**, 1–83 (2020).
64. Fuhrman, J. A., Comeau, D. E., Hagström, A. & Chan, A. M. Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. *Appl. Environ. Microbiol.* **54**, 1426–1429 (1988).
65. Giovannoni, S. J., DeLong, E. F., Schmidt, T. M. & Pace, N. R. Tangential flow filtration and preliminary phylogenetic analysis of marine picoplankton. *Appl Environ. Microbiol.* **56**, 2572–2575 (1990).
66. John, S. G. et al. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195–202 (2011).
67. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
68. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
69. De Coster, W., D’Hert, S., Schultz, D. T., Cruys, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
70. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
71. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
72. Wick, R. R. & Holt, K. E. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000 Res.* **8**, 2138 (2019).
73. Stewart, R. D. et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat. Biotechnol.* **37**, 953–961 (2019).
74. Giovannoni, S. J., Rappé, M. S., Vergin, K. L. & Adair, N. L. 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the green non-sulfur bacteria. *Proc. Natl. Acad. Sci. USA* **93**, 7979–7984 (1996).
75. R Core Team. R: a language and environment for statistical computing. <https://www.R-project.org/> (2023).
76. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
77. Quast, C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
78. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).
79. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. vol. 2nd Edition (Springer New York, New York, NY, 2009).

80. Liu, S. et al. Linkages Among Dissolved Organic Matter Export, Dissolved Metabolites, and Associated Microbial Community Structure Response in the Northwestern Sargasso Sea on a Seasonal Scale. *Front. Microbiol.* **13**, 833252 (2022).
81. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 1–13 (2021).
82. Li, H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
83. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
84. Wilke, C. Cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.3. <https://wilkelab.org/cowplot/> (2024).
85. Ihaka, R., Murrell, P., Hornik, K., Fisher, J. C. & Zeileis, A. Color-space: color space manipulation. R package version 1.3-2. <https://CRAN.R-project.org/package=colorspace> (2016).
86. Slowikowski, K. ggrepel: automatically position non-overlapping text labels with 'ggplot2'. <https://github.com/slowkow/ggrepel> (2024).
87. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
88. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 1–13 (2018).
89. Kang, D. D. et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019**, 1–13 (2019).
90. Imelfort, M. et al. GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
91. 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**, 695–607 (2015).
92. 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 Res.* **7**, 1043–1055 (2015).
93. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
94. 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 (2020).
95. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
96. Wickham, H., Pedersen, T. & Seidel, D. Scales: scale functions for visualization. R package version 1.3.0. <https://github.com/r-lib/scales> (2023).
97. Oksanen, J. et al. Package 'vegan' Title Community Ecology Package Version 2.5-7. <https://CRAN.R-project.org/package=vegan> (2020).
98. Adobe Inc. Adobe Illustrator. <https://adobe.com/products/illustrator> (2019).
99. Pebesma, E. Simple Features for R: Standardized Support for Spatial Vector. *Data. R. J.* **10**, 439–446 (2018).
100. Inkscape. Inkscape Project. <https://inkscape.org/> (2024).
101. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
102. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
103. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
104. Sullivan, M. B. et al. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* **4**, 1344–1357 (2006).
105. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
106. Buchholz, H. H. et al. Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. *ISME J.* **15**, 1585–1598 (2021).
107. Katoh, K. & Standley, D. M. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* **32**, 1933–1942 (2016).
108. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
109. 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).
110. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
111. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
112. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
113. Martínez Arbizu, P. PairwiseAdonis: pairwise multilevel comparison using adonis. R package version 0.4 (2020).
114. Cumming, G. The New Statistics: Why and How. *Psychol. Sci.* **25**, 7–29 (2014).
115. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
116. Gregory, A. C. et al. MetaPop: a pipeline for macro- and micro-diversity analyses and visualization of microbial and viral metagenome-derived populations. *Microbiome* **10**, 49 (2022).
117. Mizuno, C. M., Ghai, R. & Rodríguez-Valera, F. Evidence for meta-viromic islands in marine phages. *Front. Microbiol.* **5**, 1–10 (2014).
118. Temperton, B. BIOS-SCOPE/AE1712-viromes: v1.0.0.1. zenodo <https://doi.org/10.5281/zenodo.10940125> (2024).

Acknowledgements

The authors thank the crew and the marine technicians of the Bermuda Institute of Ocean Science vessel, the Atlantic Explorer, for the collection of seawater samples. Bioinformatic analyses were conducted using the high-performance computing resources of the Ohio Supercomputer Center provided by the Louisiana State University and those of ISCA, provided by the University of Exeter. The following grant information was disclosed by the authors: Major support was provided by a fellowship to Ben Temperton from the Bermuda Institute of Ocean Sciences as part of the BIOS-SCOPE program (Simons Foundation International); the Royal Society and the Natural Environment Research Council (NERC) (NE/PO08534/1 and NE/R010935/1). Additional support to Joanna Warwick-Dugdale from a NERC GW4+ Doctoral Training Partnership PhD (NE/L002434/1), and the BIOS-SCOPE program. Work conducted by Luis Bolanos and Rachel Parsons were funded by the BIOS-SCOPE program. The work performed by Holger Buchholz was funded by a NERC GW4+ Doctoral Training Partnership PhD. Funding support for Matthew B. Sullivan included Gordon and the Betty Moore Foundation (awards

#3790 and 5488) and the National Science Foundation (NSF) (OCE#1829831, ABI#1759874, and OCE#1829640). Support for this project also came from the NSF Center for Chemical Currencies of a Microbial Planet (C-CoMP NSF-STC 2019589). This is C-CoMP publication #041. This project utilised equipment funded by the Wellcome Trust Institutional Strategic Support Fund (WT097835MF), Wellcome Trust Multi-User Equipment Award (WT101650MA) and BBSRC LOLA award (BB/K003240/1). BT: BIOS-SCOPE (Established 2015) - Simons Foundation International; Royal Society; Natural Environment Research Council (NERC): NE/PO08534/1; NE/R010935/1. JWD: BIOS-SCOPE (Established 2015) - Simons Foundation; International Natural Environment Research Council (NERC): NE/L002434/1. RP: BIOS-SCOPE (Established 2015) - Simons Foundation. LB: BIOS-SCOPE (Established 2015) - Simons Foundation. HB: Natural Environment Research Council (NERC) GW4+ PhD. MS: Gordon and Betty Moore Foundation: 3790; Gordon and Betty Moore Foundation: 5488; National Science Foundation (NSF): OCE#1829831; National Science Foundation (NSF): OCE#2019589; National Science Foundation (NSF): ABI#1758974; National Science Foundation (NSF): OCE#1829640. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. For the purposes of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Author contributions

Joanna Warwick-Dugdale performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft. Funing Tian analysed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft. Michelle Michelsen performed the experiments. Dylan R Cronin analysed the data. Karen Moore performed the experiments, contributed reagents/materials/analysis tools. Audrey Farbos performed the experiments. Lauren Chittick performed the experiments. Ashley Bell analysed the data and prepared figures. Holger Buchholz analysed the data. Ahmed A Zayed analysed the data. Luis Bolanos-Avellaneda analysed the data and prepared figures. Rachel Parsons contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft. Michael J Allen authored or reviewed drafts of the paper, approved the final draft. Matthew B Sullivan contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.

Ben Temperton conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-48300-6>.

Correspondence and requests for materials should be addressed to Joanna Warwick-Dugdale or Ben Temperton.

Peer review information *Nature Communications* thanks Cynthia Silveira and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024