Long-read metagenomics reveals cryptic and abundant marine viruses

Authors
Joanna Warwick-Dugdale[1,2], Natalie Solonenko[3], Karen Moore[2], Lauren Chittick[3], Ann C. Gregory[3], Michael J. Allen[1,2], Matthew B. Sullivan[3,4], Ben Temperton[*2].

Addresses
[1.] Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth, PL1 3DH, UK. [2.] School of Biosciences, University of Exeter, Stocker Road, Exeter, EX4 4QD, UK. [3.] Department of Microbiology, and [4] Civil, Environmental and Geodetic Engineering, The Ohio State University, 105 Biological Sciences Building, 484 West 12th Avenue Columbus, Ohio 43210, USA.

*Corresponding author; email: b.temperton@exeter.ac.uk

Introduction

Marine bacteriophages impact global biogeochemical cycles via their influence on host community structure and function[1], yet our understanding of viral ecology is constrained by limitations in culturing of important hosts and the lack of a 'universal' gene to facilitate community surveys[2]. While recent advances in short-read viral metagenomics offer solutions[2–6], they are confounded by microdiversity issues[4,7] and fail to assemble over genomic islands (GIs)[8,9]. Single-virus genomics can overcome such issues by targeted sequencing of single virus particles[4], but is technically challenging and costly to implement. Another approach would be to leverage long-read sequencing technologies to bridge GIs and connect SNPs within populations, but current technologies require far too much input material (micrograms)[10] and have high error rates[11]. Here we sought to establish a generalizable, long-read, low-input metagenomic sequencing approach (VirION) to survey viruses in nature and applied it to nanogram concentrations of DNA from mock and natural viral communities. These experiments showed that our VirION method (i) is as relatively quantitative as short-read methods, (ii) captured many abundant and ubiquitous viral genomes that short reads did not, (iii) significantly increased median contig lengths and captured a complete 316 kbp viral genome - 100 kbp longer than the current longest genome from Global Ocean Viromes (GOV), (iv) overcame issues of microdiversity in viral assemblies and (v) captured more genomic islands than short-read assemblies, providing insight into the gene content of whole genomic islands from metagenomic data. Thus, long-read viral metagenomics adds another tool to the maturing viral survey toolkit, recovering ubiquitous and abundant viral taxa missed by short-read sequencing.

The marine bacterial communities that regulate global biogeochemical cycles are themselves structured by selective, phage-mediated lysis[12]. Bacteria co-evolve with their phages and exchange genetic information[13], and phages even 'reprogram' hosts during infection[14], channelling host metabolism towards processes involved in phage replication[15]. Recent advances in viral metagenomics have provided clues as to viral function (e.g. virally encoded, host-derived central metabolism genes (Auxilary Metabolic Genes: AMGs))[15,16], and first estimates of global viral gene abundance and distribution[3,17]. Yet short-read assemblies

are composites of population 'features'[8], with successful assembly a function of coverage and branch-resolution in assembly graphs[18,19]. Genomic regions of high diversity, such as GIs, lack sufficient coverage or are bounded by repeat regions[9], and therefore fail to assemble using current De Brujin Graph (DBG) assembly methods. Even in extremely abundant viruses, such as the pelagiphage HTVC010P[20], GIs and/or high microdiversity may cause genome fragmentation and subsequent removal from downstream analyses[4,7]. Fosmid-based[5,6] and single-virus genomics[4] approaches can overcome short-read assembly failure associated with these issues, but are expensive and time-consuming.

A significant obstacle in adopting long-read technology for marine metagenomics lies in obtaining the amount of DNA required: Viral DNA extraction from 20 L of seawater yields far less than the micrograms of DNA recommended for efficient long-read sequencing[10]. To overcome this limitation, we developed a Long-Read Linker-Amplified Shotgun Library approach for long-read viral metagenomics (named VirION; Figure 1), using 20 ng of viral DNA randomly sheared to ~ 8 kbp fragments and amplified by polymerase chain reaction (PCR) for input into a ligation sequencing protocol on the Oxford Nanopore MinION. VirION was evaluated using a mock viral community comprising six phages from the order Caudovirales (Supplementary Table 1). In total, 348,317 high quality reads (median length: 4,142 bp; max length: 24,940 bp) were generated from a single flow-cell. PCR product length was observed as reduced (~5.4 kbp) compared to template length, indicating PCR preference for amplification of shorter DNA fragments (Supplementary Figure 1a). Further reduction in average read length (Supplementary Figure 1b) may result from preferential diffusion (and thus sequencing) of shorter reads within the flowcell.

Assembly of long reads recovered all six viruses at >98.7% completion on 22 contigs (>500 bp), with an N75 of 18.5 kbp. In comparison, a matching short-read assembly of 5.6M 2 × 300bp reads yielded 579 contigs, with an N75 of 0.6 kbp. Median coverage of mock viral genomes and mapped reads per genome were strongly correlated between VirION reads and short-read data (R^2=0.975, $p<0.001$for coverage and mapped reads), showing long-range amplification did not introduce additional bias to relative abundance (Figure 2a). Error-rates in long-read data were significantly higher (6.4%) than short read data (0.9%). Consequently, long-read assemblies contained 1,787 base-calling errors (identified as indels upon alignment to reference genomes) across the six members of the mock community, compared to 12 for short-read assemblies. However, correction of long-read assemblies with the matching short read dataset by consensus base-calling successfully corrected 1,729 (~97%) of the base-calling errors. 9,620 out of 359,338 mapped long reads (2.6%) were classified as chimeric. Thus, limiting PCR to 15 cycles was sufficient to avoid high numbers of chimeric sequences in this study, compared to rates of 6.5% with 29 cycles of PCR observed in previous studies[21]. Contigs assembled from long reads using the Overlap Layout Consensus (OLC) assembler, Canu[22] showed no chimeric assemblies when aligned to genomes from the mock community.

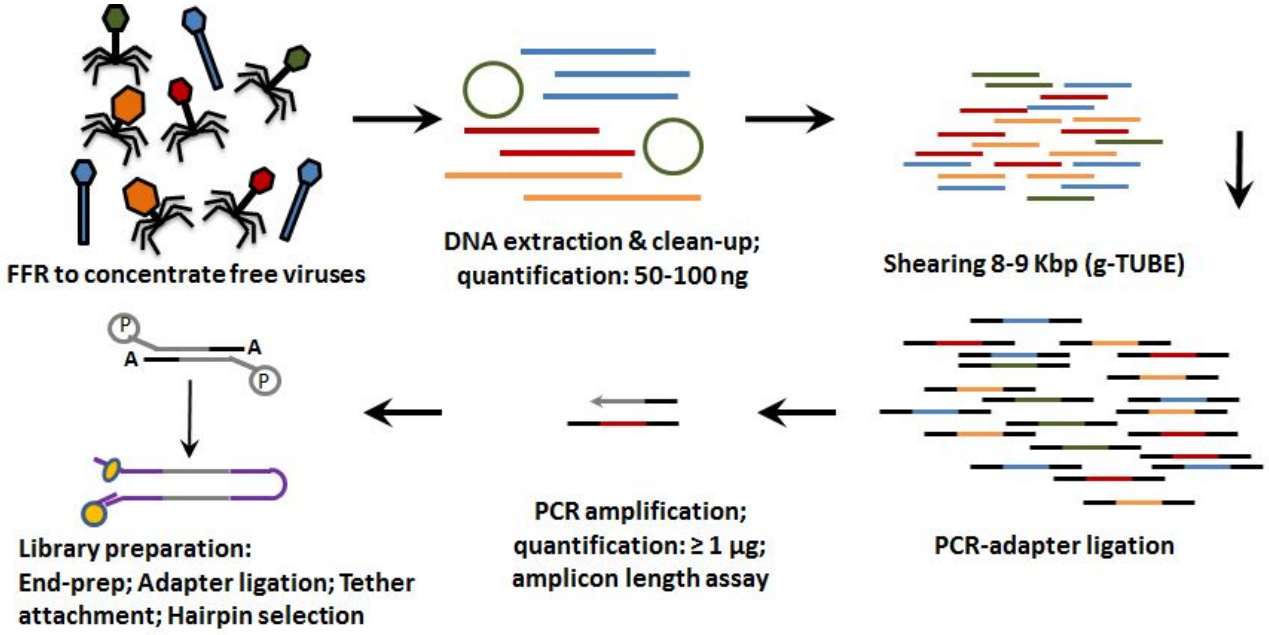

Figure 1: Basic workflow for preparation of free-viral fraction DNA for MinION sequencing. The long-read viral metagenomic method developed includes $FeCl_3$ flocculation and resuspension (FFR), shearing of extracted viral DNA (to 8-9 kbp), random linker amplification (Linker Amplified Shotgun Library: LASL), MinION library preparation, and nanopore (Oxford Nanopore Technologies; ONT) sequencing.

Following optimisation of library construction and amplification parameters, we applied VirION to a natural viral community collected from the Western English Channel (WEC) coastal station 'L4' to evaluate whether optimised long read viral metagenomics provided new insights into viral communities. A single MinION flow-cell of VirION sequencing produced 108,718 high quality reads (median length: 3,625 bp; max length: 17,019 bp) with a total yield of 0.39 Gbp. In addition, a matching short-read library was deeply sequenced (58.6M 2 × 300 bp). To combine long and short reads for maximal gain, a viral metagenomic bioinformatic pipeline was developed (Supplementary Figure 2). Briefly, short (Illumina) and long (MinION) sequences were assembled separately to generate short-read only contigs (via DBG) and long-read only contigs (via OLC). Short and long reads were also combined in a hybrid metagenomic assembly which uses long reads to scaffold short-read (DBG-derived) contigs. Next, DBG short-read and hybrid scaffolds were combined and clustered using average nucleotide identity (producing ANI clusters), to de-replicate near-identical viruses into phage OTUs[23] from short-read only and hybrid DBG approaches. Error corrected contigs > 10 kbp from long-read only assemblies (OLC) were pooled with DBG cluster representatives, and putative viral genomes were identified using VirSorter[24]. Putative viral contigs were organised into viral clusters (VCs) based on encoded proteins using vContact2[25] and whole-genome nucleotide diversity[26] of DBG and OLC contigs was evaluated to assess the efficacy of using long-reads to assemble microdiverse viral genomes. In total, our analyses generated 2,645 putative viral contigs from the WEC dataset (2,279 from DBG (combined short and hybrid) assemblies and 366 from error-corrected OLC assemblies).

First, we investigated the effect of sequencing depth on the efficiency of short-read and hybrid approaches. Below or at the mean sequencing depth achieved in the GOV dataset (~9 Gbp), hybrid assemblies of triplicated subsamples captured significantly more (between 1.1 to 1.5-fold, Student t-test, $p<0.05$) putative viral genomes >10 kbp than short-read assemblies . However, at higher sequencing depths, inclusion of long reads did not significantly increase the number of viral genomes recovered (Figure 2b). At sequencing depths equal to, or greater than the mean sequencing depth of the GOV dataset, hybrid assemblies generated significantly more (1.5 to 2.0-fold, Student t-test, $p<0.05$) circular (i.e. putatively complete) viral genomes than short-read assemblies (Figure 2b). These results indicate that when samples are sequenced to extreme depth, hybrid assembly offers no advantages in identifying more partial viral genomes, but that the hybrid approach proved superior in capturing more complete viral genomes, even when high levels of Illumina coverage were applied.

Next, we wanted to determine whether long reads could significantly increase contig length. Alignments within ANI clusters showed that long reads can be used to scaffold short-read assemblies to produce longer contigs in viral metagenomic datasets (Figure 2c). The longest complete viral genome was a 316 kbp hybrid assembly (H_NODE_6) that joined two short read contigs (204 kb and 112 kb, respectively). This genome was 100 kbp longer than the longest genome from the GOV(GOV_bin_1520_contig-100_0) and was found in a VC with several of the largest phage genomes ever sequenced in RefSeq, including Myoviridae *Cronobacter sakasakii* phage GAP32 and Enterobacter phage vB_KleM-RaK2, at 359 kb and 346 kb respectively[27,28], demonstrating the capacity for hybrid assemblies to fully sequence very large phage genomes from metagenomic data. As suggested by these results, the median length of hybrid ANI cluster representatives was found to be significantly longer (1,769 bp) than that of short-read assembled representatives (Mann-Whitney U test, n_1=1400, n_2=879, p-value<0.001).

We also evaluated the use and amount of short-read data required to correct basecalling errors in long-read environmental metagenomic data. 3-27.8 Gbp of short read data from the WEC was subsampled in triplicate from the full dataset and mapped against OLC contigs for error-correction with Pilon. At ~9 Gbp (median coverage of ~70), the limit of fixed deletions and insertions was achieved (Supplementary Figure 3). Thus, use of the full dataset for Pilon error-correction in this study is likely to have fixed as many errors as possible. Indels resulting from base-calling errors typically produce shorter predicted coding sequences than error-free assemblies via the introduction of stop codons. We compared predicted protein length distributions in viral contigs before and after error correction, to predicted proteins derived from (a) DBG assemblies (both short and hybrid) from the WEC and; (b) genomes of *Caudovirales* isolates from the RefSeq v8.4 viral database. As expected, error-prone raw long reads resulted in an abundance of short predicted proteins (Figure 2d). Error correction during OLC assembly provided a modest improvement in predicted protein length, but still below that of current DBG assembly of short-reads. Polishing of long read OLC contigs with short read data restored predicted protein lengths to those of isolated *Caudovirales* genomes.

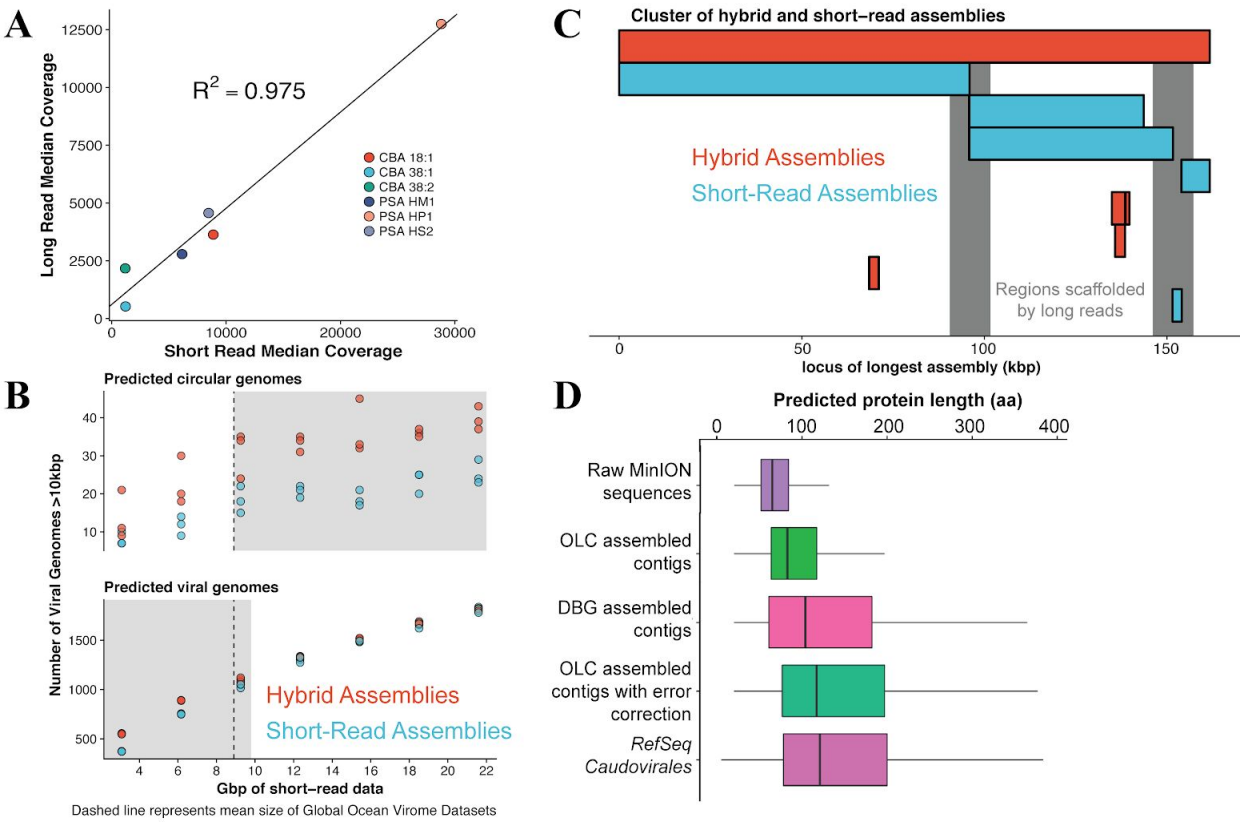

Figure 2: **Comparative performances of short-read and long-read data for the identification of marine viral genomes. a,** Relative abundances of genome-mapped MinION sequences (~4 kbp) and short-read Illumina assemblies from a mock viral community composed of 6 different tailed bacteriophages. CBA: Cellulophaga phage; PSA: Pseudoalteromonas phage. The relative abundances of mock viral community members were approximately equal for short-read and long-read data, as shown by their proximity to the dotted identity line. Therefore, the random linker amplification (Linker Amplified Shotgun Library: LASL) method employed in the generation of sufficient amounts of viral DNA for MinION sequencing, and the sequencing itself, was as quantitative as Illumina sequencing for estimating relative viral abundance**. b,** Efficiency of short-read only and hybrid sequencing approaches for detection of viral genomes at various depths/coverages of Illumina data using triplicate pseudorandom subsamples of short read data from the Western English Channel (WEC) viral metagenomes: At all coverage depths tested, hybrid assemblies generated more circular (i.e. putatively complete) viral genomes than short-read assemblies; Below 10 Gbp of short-read data, hybrid assemblies captured more viral genomes (> 10 kbp) than short-read assemblies. Comparisons within grey boxes were found to be statistically significant (Student t-test). **c,** An example of long reads scaffolding short-read assemblies in putative viral contigs using a DeBrujin Graph approach. Here, the blue contigs from short-read only assembly align to the longer hybrid assembly contig (red) at >95% identity over 80% of their lengths. Breaks in the short-read only contigs can be clearly seen and are highlighted here in grey. **d,** Distributions of predicted viral protein length for raw MinION reads; OLC assembled MinION reads before and after error-correction with matching short read data from the WEC viral metagenomes; DBG assemblies (hybrid and short-read only), and *Caudovirales* isolate genomes from RefSeq. Error-prone MinION reads contain indels that reduce the length of predicted proteins relative to those observed in both WEC hybrid contigs, and tailed phage genomes from the RefSeq viral database v8.4. Error-correction of long-read contigs with short-read data restored predicted protein length in long-read contigs to match those of viral isolates.

It has been hypothesized that genomes assembled from short-read metagenomes may be biased away from microdiverse populations[4,7]. We reasoned that error-corrected long-read metagenomes might better capture such microdiverse populations. To test this we assessed genome-level nucleotide diversity[26] (π) of both DBG and OLC-derived contigs. Median nucleotide diversity in OLC-assembled contigs was

significantly (3-fold) higher than those derived from DBG assemblies (two-sided Mann-Whitney U test: W=105,830, n_1= 758, n_2=206, $p = 4.81 \times 10^{-15}$) (Supplementary Figure 4). This result supports the hypothesis that OLC assembly of long reads can recover genomes previously lost due to failure to resolve assembly graphs as a consequence of microdiversity. We also reasoned that long-reads would be able to span putative GIs to successfully capture genome content from metagenomic data, potentially giving greater insight into host-virus interactions and the proteins involved in phage replication. A GI was defined as a region >500 bp in length with median coverage of <20% of the median coverage of the whole contig[8]. In total, we identified 1.65-fold more GIs per kbp in OLC-derived contigs than in DBG-derived contigs (0.19 versus 0.12, respectively, p<0.0001, Supplementary Figure 5a). GIs captured on OLC-derived contigs were also 1.76-fold longer compared to DBG-derived contigs (254 bp versus 145 bp per kbp of contig, respectively; p<0.0001; Supplementary Figure 4b).

An excellent example of a GI captured by this study is provided by 'tig404', an OLC-derived contig closely related to pelagiphage HTVC010P (a highly abundant, cultured podovirus seldom recovered from metagenomic assembly due to high microdiversity). A DBG assembly of short reads mapping to tig404 resulted in 2,005 scaffolds, with a maximum contig length of 3,527 bp, which is below the 5 kbp cut-off usually recommended for robust identification of viral contigs[24,29] (Figure 3a). Conversely, an OLC assembly of mapped long reads resulted in a single contig 28.9 kbp in length (Figure 3b). This single contig shared a 5.3 kbp GI similar to those previously observed on the cultured isolates of HTVC010P and HTVC010P-like sequences from an environmental fosmid library[5]. Furthermore, it was observed that mapping of unassembled long reads to the boundary regions of known GIs from cultured isolates, and from short-read assemblies, provided insight into their genomic content at the population level. In HTVC010P, long reads mapping to the GI boundary revealed the conserved pattern of a ribonuclease upstream of an internal virion protein, putatively used by T7-like phages to penetrate the host membrane[8] (Figure 3c). Therefore, long-read mapping showed that, at least in these HTVC10P-like sequences, the functional capacity of the GI is conserved and is likely associated with the host-virus co-evolutionary arms race.

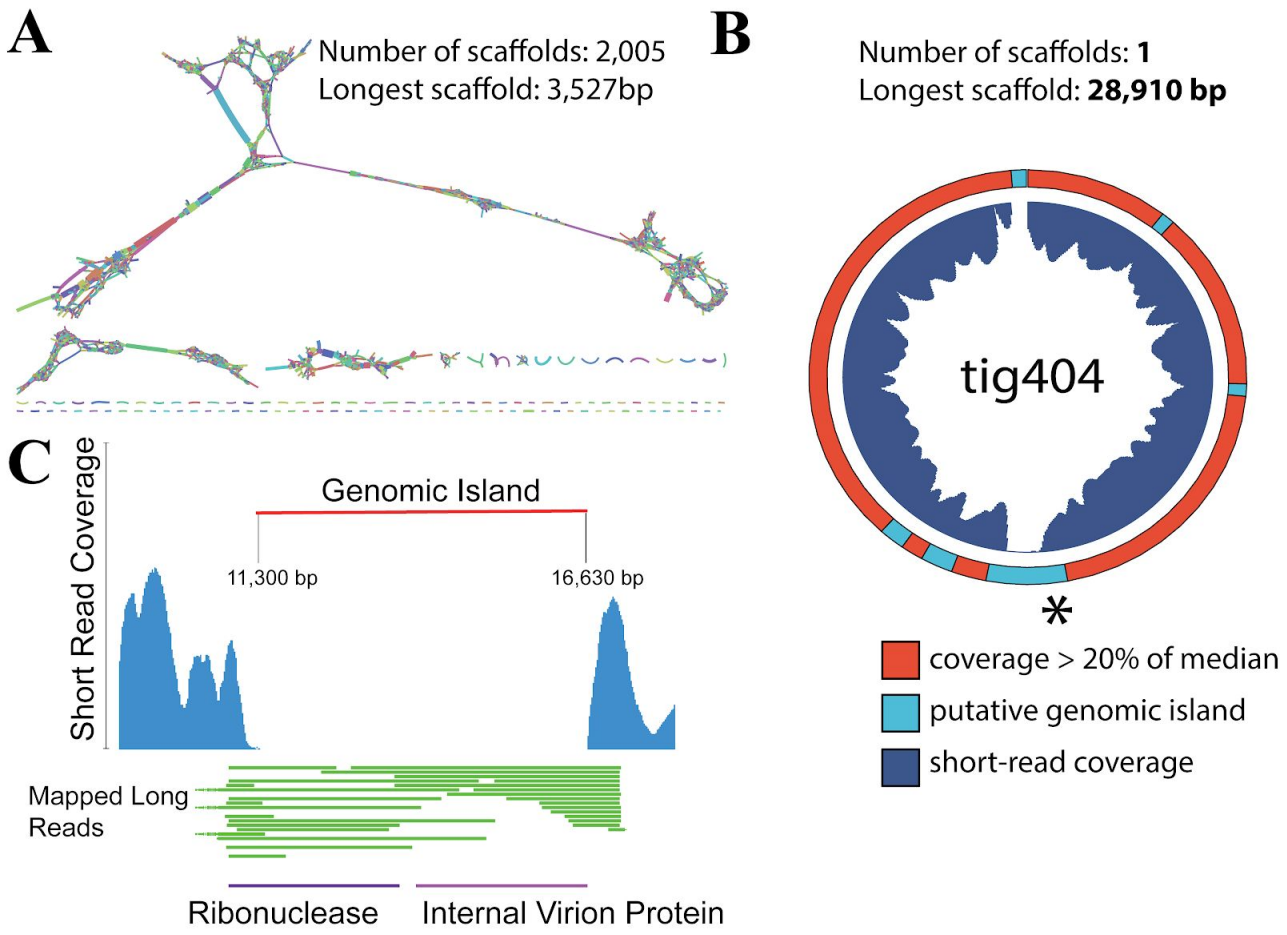

Figure 3: Long-read sequencing resolves microdiversity and abundance related assembly issues and provides insight into genomic islands in metagenomic data. A, Highly similar strains within a viral community result in unresolved branching during De Brujin Graph assembly (DBG), here demonstrated in this Bandage visualisation of a short-read assembly that maps to a novel Western English Channel virus ('tig404') closely related to pelagiphage HTVC010P; **B,** For the same novel Western English Channel virus ('tig404'), Overlap consensus assembly (OLC) of long reads has resolved branching issues, producing a single scaffold of over 28 kbp in length. Short-read coverage (dark blue) identified six putative genomic islands (light blue) successfully assembled by the long reads, including one (*) that encoded a ribonuclease and an internal virion protein (thought to be involved in host-virus interaction, and similar to that observed in HTVC010P). **C** Mapping of long reads against contigs and genomes of cultured isolates, such as HTVC010P, provides insight into population-level genomic diversity within genomic islands by extending sequence space from conserved borders into the island.

Next, we wanted to assess the abundance of our variously-derived viral genomes. We calculated the relative abundance of DBG and OLC-derived contigs in the WEC by competitive mapping of short reads. 60% of the top 50 most abundant viruses in the WEC sample were derived from OLC assemblies (Supplementary Figure 6). We then wondered whether the long-read assembly contigs were abundant in short-read datasets, but had been missed in previous studies due to microdiversity and/or GIs. To test this, we pooled DBG and OLC-derived contigs >10kbp with those from the GOV and other viral datasets (Supplementary Table 2), and evaluated relative-abundance by competitive mapping of a subsample of 10 million reads from 42 GOV surface datasets (Figure 4). Of the top 50 most abundant viral genomes

identified across all GOV surface samples, 22 (34%) were viruses obtained from the WEC via OLC assembly, and 7 (8%) were derived from DBG hybrid assemblies (WEC short-read only contigs comprised a further 6%) (Supplementary Figure 7).

The most globally abundant and ubiquitous viral genome was a DBG-hybrid assembled contig (22.4 kbp in length) from our study, called 'H_NODE_1248' (4th most abundant across all samples, regardless of ubiquity). H_NODE_1248 occupied a VC with 57 other members, including vSAG-37-F6 (9th most abundant ubiquitous virus and 13th most abundant across all samples), previously identified as putatively the most globally abundant virus[4], and 10 OLC-assembled contigs ranging in size from 10 kbp to 27 kbp. Interestingly, pelagiphage HTVC010P, once thought to be the most abundant virus on Earth[20], was ranked 128th in global abundance and failed to match the criteria of being both ubiquitous and abundant. This finding supports caution when claiming high abundance of novel phages, as novelty can inflate abundance estimates calculated by read alignment[20]. As new similar sequences are discovered and the novelty is diminished, reads will be divided across multiple sequences and the abundance of an individual contig will decrease (an inevitable conclusion of the increasing resolution of viral sequence space from metagenomic studies). The importance of these results are clear: Over half of the most abundant WEC viruses identified in this study, a major component of inter-cluster viral diversity across global oceans, would not have been identified as such without the inclusion of error-corrected, long-read sequences in the viral metagenomic pipeline. Do any of these new OLC-derived contigs represent novel viral types? In total, clustering by shared protein content produced 668 VCs, which are approximately equivalent to ICTV-accepted viral genera[25]. Of these, 202 VCs contained OLC-derived contigs, but just 3 of these were comprised solely of OLC-derived contigs, suggesting that OLC-derived contigs do not dramatically increase the total number of VCs observed in marine surface viral metagenomes. This supports previous findings that viral diversity at the genera level in surface oceans has been sufficiently documented[3].

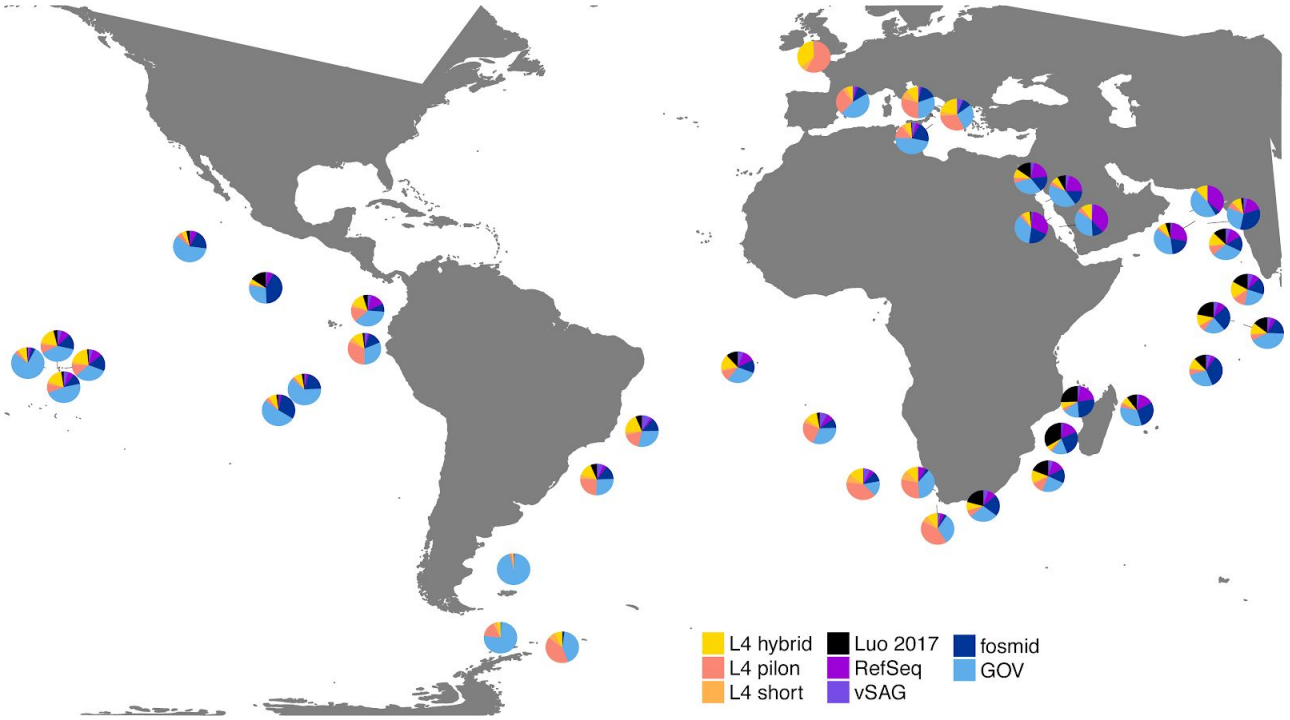

Figure 4: Globally distributed relative abundances of WEC viruses across Global Ocean Virome (GOV) samples. Abundances are calculated via competitive recruitment of 10 million sub-sampled GOV short reads against: Viral genomes (circular and fragmented) obtained in this study (from short-read, hybrid, and error-corrected long-read viral assemblies); Viral genomes obtained from other key viral metagenomic studies (including those which have employed short-read sequencing ('GOV'[3]; 'Luo 2017'[30]), and long-sequence recovery via Single-virus genomics ('vSAG'[4]), and fosmid libraries ('fosmid'[5,6])), and cultured viruses (from the NCBI RefSeq database) (all detailed in Supplementary Table 2).

In summary, this investigation represents the first use of long-read sequencing for viral metagenomics. We have shown that combining long-read sequencing with short reads, in both hybrid assemblies and for error correction of long-read assemblies, successfully overcomes issues of genome fragmentation in microdiverse, ubiquitous and ecologically important viral taxa across a broad range of viral clusters. Furthermore, long read sequencing enables greater resolution of the content of hypervariable genomic islands associated with host-virus interactions, providing an additional tool for improving sequence-based viral ecology. We have also shown that phage genomes captured from the Western English Channel via these techniques are ubiquitous across global oceans, and currently absent from oceanic datasets. Importantly, long-read sequencing on the MinION platform is undergoing rapid improvements in terms of yield, with current technology providing at least an order of magnitude more sequencing data than that produced in this study. In conjunction with the protocols developed here, long-read sequencing of viral metagenomes offers the potential to significantly improve our understanding of viral ecology in global oceans, human microbiomes and agriculture.

Methods

Mock viral community: A mock viral community composing of six different marine Caudovirales from the three families (Myoviridae; Podoviridae; Siphoviridae), with genome sizes ranging from 38.5-129.4 kbp, and GC contents from 33-45% (Supplementary Table 1), was produced as described previously[31]. Briefly, viruses were cultivated from host *Pseudoalteromonas* or *Cellulophaga* via plaque assay, collected into MSM buffer (0.45 M NaCl, 0.05 M Mg, 0.05 M Tris base, pH 7.6) and purified by 0.2µm filtration followed by treatment with DNase I (100 U/mL for 2 hr at RT; terminated by the addition of 0.1 M EGTA and 0.1 M EDTA). Viral capsids were enumerated via epifluorescence microscopy (SYBR Gold; wet mount method)[32,33] . 1.4×10^9 virus particles from each culture were pooled, and DNA extracted via the Wizard® DNA Clean-up System (Promega A7280). DNA was quantified via Qubit fluorometer (Life Technologies).

L4 virome: Approximately 20 L of seawater was collected in rosette-mounted Niskin bottles at a depth of 5m from the Western Channel Observatory (WCO; http://www.westernchannelobservatory.org.uk/) coastal station 'L4' (50°15.0'N; 4°13.0'W) on the 28th September 2016. Seawater was transferred immediately to a clean collection bottle, and processed to remove the cellular fraction (within 4 hours of collection) via sequential filtration through glass fibre (GF/D: pore size 2.7 µm) and Polyethersulfone (PES: pore size 0.22 µm) filters in a 142 mm polycarbonate (PC) rig, with peristaltic pump. Precipitation of viruses from filtrate (i.e. the viral fraction) and primary concentration of virus particles was conducted by iron chloride flocculation and collection on large-pore (0.8 µm, PC) filters[34] and stored in the dark at 4°C.

Viruses were resuspended in ascorbate-EDTA buffer (0.1 M EDTA, 0.2 M $MgCl_2$, 0.2 M ascorbic acid, pH 6.1), and transferred to Amicon Ultra 100 kDa centrifugal filter units (Millipore UFC910024)[35], that had been pre-treated with 1 % bovine serum albumin buffer to minimise capsid-filter adhesion[36] and flushed with SM buffer (0.1 M NaCl; 0.05 M Tris-HCl; 0.008 M $MgCl_2$). Following concentration to 500-600 µL, virus particles were washed with SM buffer[37] and purified with DNase I (100 U/mL; 2 hr at RT) to remove unprotected DNA (i.e. encapsulated DNA); DNase I activity was terminated by the addition of 0.1 M EGTA and 0.1 M EDTA[35]. Viral DNA was extracted from concentrated and purified viral particles using the Wizard® DNA Clean-up System (Promega A7280), removing PCR inhibitors (e.g. EDTA)[34].

Library preparation, amplification, and sequencing: For short-read sequencing, Illumina libraries were generated from 1 ng of mock viral community DNA, and 1 ng of environmental viral-fraction DNA, using Nextera XT v2 kits (Illumina) and the manufacturer's protocol. After 12 cycles of amplification, the concentration and distribution in fragment sizes of the Illumina libraries were determined via Qubit and Bioanalyzer (Agilent), respectively. 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); 60 million paired-end reads were produced. For long-read (MinION) sequencing, 20 ng (mock viral community)

or 100 ng (WEC viral-fraction) of DNA was sheared to fragments averaging 8 kbp length via g-TUBE (Covaris 520079), as required to optimise MinION flow cell sequencing efficiency/yield (Oxford Nanopore Technologies: ONT). For end-repair of DNA fragments, amplification of DNA with PCR-adapter ligation (i.e. Linker Amplified Shotgun Library: LASL preparation), and preparation of MinION-compatible libraries, the 'Ligation Sequencing kit 2D' (ONT SQK-LSK208), was employed, following the manufacturer's protocols for "2D Low input genomic DNA with PCR", with some modifications to improve the sequencing of viruses. Briefly, the following steps were undertaken: 1) DNA fragmentation (as above); 2) End-repair, dA-tailing (NEBNext End Repair Module (NEB E7546S)) and 1 × AMPure bead clean-up (note that the control sequence derived from lambda phage was omitted to prevent issues with downstream analysis of viral sequences); 3) PCR adapter ligation (PCR adapters (ONT), NEB Blunt / TA Ligase Master Mix (NEB M0367S)) and 0.4 × AMPure bead clean-up; 4) PCR amplification (using 20 ng of PCR-adapter ligated viral DNA per reaction, NEBNext High-Fidelity 2× PCR Master Mix (NEB M0541S), primers (matching PCR adapters: ONT); Note that PCR reaction conditions were modified with reference to NEBNext manufacturer's instructions in order to maximise DNA yield, whilst minimising production of chimeric sequences, as follows: 3 mins at 95°C (initial denaturation), 15 cycles of: 15 secs at 95°C (denaturation), 15 secs at 62°C (annealing), 5 min at 72°C (extension); finally 5 min at 72°C (final extension)) followed by 0.4 × AMPure bead clean-up); 5) Second end-repair, dA-tailing and AMPure bead clean-up (as step 2, but using ~1.5 µg of amplified DNA, and a 0.4 × bead wash to minimise carryover of short amplification products); 6) Sequencing-adapter ligation (including hairpin adapter ligation and tethering (ONT)); 7) Purification of adapted DNA (MyOne C1 Streptavidin beads (Thermo Fisher Scientific Inc. 65001)); 8) Elution of library (elution buffer: ONT). A 1D version of this protocol using current Nanopore chemistry (with no Streptavidin bead purification) has been tested and is available on protocols.io here. All ethanol washes conducted at 80% EtOH with nuclease-free water (NFW); elution from AMPure beads at 55°C to maximise recovery of long DNA fragments (elution from C1 beads conducted at RT). The MinION-compatible libraries were sequenced on a MinION Mk 1B (ONT) with R9.4 flow cells. Libraries were loaded as per ONT instructions using Library Binding Beads. Base-calling of the data collected during sequencing was performed using ONT software, either in real-time (using the cloud-based program 'Metrichor'), or after sequencing (with the locally installed program 'Albacore').

Generation of short-read and hybrid assemblies: Following the removal of adapters and quality filtering with Trimmomatic[38], high quality short-read sequences were normalised and error-corrected with bbnorm (https://github.com/BioInfoTools/BBMap) before assembly with metaSPAdes[39] DeBrujin Graph Assembler (with error-correction disabled). Long reads were processed with NanoFilt (https://github.com/wdecoster/nanofilt) to keep sequences with a q-score > 10 and the first 50 bases were removed to increase sequence quality, with a minimum sequence length of 1 kb. Sequences were checked for adapters using Porechop (https://github.com/rrwick/Porechop). OLC assembly was performed using Canu[22] with the following parameters: "genomeSize=180m minReadLength=1000 contigFilter="2 1000 1.0 1.0 2"

corOutCoverage=999 correctedErrorRate=0.040 -nanopore-raw". genomeSize parameter selection was optimised by testing values from 9-180 Mbp on mock viral community data and evaluated using metaQUAST[40]. Above a value of 45 Mbp, all assemblies were nearly identical, so the largest value was chosen for downstream analyses. Canu assemblies were polished using Pilon[41] v1.22 with short-read sequences from either the mock viral community or the WEC dataset. Comparisons of assembly methods of mock viral communities, including presence of indels were evaluated using metaQUAST[40]. Following DBG assembly of WEC sequence data, scaffolds >2.5 kbp were de-replicated using a cutoff of 95% average nucleotide identity (ANI) over 80% of the length (via MUMmer[42] v3.23) to cluster highly similar contigs from hybrid and short-read only assemblies. The longest representatives of each clusters were carried forward for analysis. Polished long read contigs or DBG cluster representatives >10 kbp from the WEC sample were processed with VirSorter[24] (in virome decontamination mode) to identify putative viral contigs (classified as category 1,2,4 or 5).

Relative abundance of VirION sequences on mock viral community and effect of sequencing depth on genome recovery in hybrid assemblies: Relative abundance of mock viral community members in short-read and VirION datasets was evaluated by mapping high quality short reads and long reads against the genomes of mock viral community members using bowtie2[43] and minimap2 (https://github.com/lh3/minimap2), respectively. Quantification of error rates in short and long read sequences were calculated with 'samtools stats'. Chimeric raw reads were identified as those that had alignments > 100 bp to two or more viral genomes, or those that had two or more alignments to the same viral genome, but did not align to the start of the genome (to avoid breakage of mappings across a linear representation of a circular genome). Mock community assemblies were evaluated for chimeric assemblies using MUMmer[42] against member genomes and identifying those that aligned to more than one member. To assess the impact of short-read sequencing depth on hybrid assemblies, high-quality short read sequences were subsampled to between 10% and 70% of the full dataset in triplicate using seqtk (https://github.com/lh3/seqtk), followed by assembly with metaSPAdes (REF) with or without support from MinION reads. Scaffolds >10 kbp in replicated assemblies were uploaded to VirSorter[24] for classification using virome decontamination mode. Numbers of scaffolds classified as viral (category 1,2,4 or 5) were calculated for each replicated assembly. Statistical significance of the number of viral or circular viral contigs between hybrid and short-read assemblies was calculated by a two-sided Student t-test between replicates for each subsample. Details of differences in means and p-values are available in Supplementary Tables 3&4.

Evaluating the nucleotide diversity of OLC vs. DBG contigs: High quality short reads from the WEC were mapped back to viral contigs using bowtie2 and nucleotide diversity was calculated using MetaPop. Briefly, short reads from the WEC dataset were mapped back against DBG and OLC-derived contigs. Reads mapping at <95% identity to any viral contig were removed, as are contigs with <10-fold coverage across

70% of their whole genome. Single nucleotide polymorphisms (SNPs) were identified using mpileup and BCFtools(https://samtools.github.io/bcftools/bcftools.html) and those with a quality score >=30, represented by at least 4 reads and comprising >1% of the base pair coverage for that position were considered true SNPs. SNP frequencies across all genomes were rarefied by subsampling to 10× coverage proportionate to the frequency of different SNPs per site while maintaining SNPs linkages. Observed nucleotide diversity (π) was estimated per contig.

Identification of GIs in WEC contigs: We created a pool of viral contigs containing 1) The viral genomes >10 kbp identified in this study; 2) A selection of phage genomes >10 kb from key metagenomic studies[3,4,30] (Supplementary Table 2); 3) Cultured viruses from the NCBI RefSeq viral database (v8.4). High quality short-read sequences were mapped against this pool using bbmap (https://github.com/BioInfoTools/BBMap) with the following parameters: 'ambiguous=all idfilter=0.90 (or 0.98) lengthtag=t'. The resulting bam file was filtered as performed in a previous study of GIs in viral genomes[8] with one modification - a minimum identity of 90% (rather than 98%) was applied for read alignment to account for higher error rate in uncorrected long read sequences. Contigs with a median coverage of >5 and a Reads Per Kilobase of contig per Gigabase of short-reads (RPKG) value of > 1 were selected for downstream analysis. To identify GIs, a sliding window analysis was performed on each contig to identify any regions >500 bp where the median coverage of the window was <20% of the median coverage of the contig . Each GI was defined as the maximum length of windows that met this criteria. The total number of GIs per contig, number per kb of contig, total length per contig were calculated for OLC-derived contigs and DBG-derived contigs from the WEC. Enrichment of GI number and length in OLC-derived contigs was evaluated by comparing measured values for OLC-derived contigs to a distribution of values from 10,000 random samples (without replacement) of an equal number of contigs.

Estimating relative abundance and viral clusters of WEC viruses in viral metagenomes: FastViromeExplorer[44] v.1.1 was used to quantify the relative abundances of WEC viral contigs. FastViromeExplorer is built upon the Kallisto[45] framework and competitively recruits reads against contigs, allowing for accurate recruitment to contigs that may share a degree of sequence similarity. Briefly, high quality short read datasets from the GOV[3] and from our WEC sample were subsampled to 10 million reads using seqtk, and recruited against a Kallisto index comprising 1) The viral genomes >10 kbp identified in this study; 2) A selection of phage genomes >10 kb from key metagenomic studies[3,4,30]; 3) Cultured viruses from the NCBI RefSeq viral database (v8.4). To ensure uniformity of gene-calling (for vCONTACT2) and viral classification all phage genomes were re-run through VirSorter (v.1.03) on the CyVerse cyberinfrastructure[46]. A total of 17,891 contigs were identified as viral and within categories 1, 2, 4, or 5 (Supplementary Table 2). For inclusion in downstream abundance analyses, contigs with less than 40% coverage as calculated by FastViromeExplorer were classified as having zero abundance. The top 100 most abundant contigs from each sample were selected for downstream analyses.

To assess whether long-read sequencing increased current estimates of global viral diversity, all sequences > 10 kbp and identified as viral by VirSorter were processed with vContact2 on the CyVerse infrastructure. In the final stage of clustering, VContact2 uses ClusterONE[47] and assigns a p-value to a cluster depending on whether the in-cluster edge weights are significantly higher than the out-cluster edge weights. Q-values were calculated from cluster p-values using the qvalue R package[48] to account for multiple testing and a q-value cutoff of 0.05 was used to identify statistically significant clusters. Additionally, clustering was used to evaluate error-correction of OLC reads. Assuming that uncorrected indels would reduce the similarity of protein coding between two contigs due to the introduction of stop codons, error-prone contigs should be more weakly connected to their in-cluster neighbours than DBG contigs. To test this, clusters with a q-value of <0.05 that contained at least 10 members and at least 1 OLC member were identified. In-cluster edges were labeled as either 'GOV_GOV', 'GOV_L4_pilon', 'GOV_L4_short', 'GOV_L4_hybrid', 'GOV_RefSeq', depending on the target and the source vertices. A one-way Mann-Whitney U test was then calculated on each label (with appropriate q-value correction) to determine if connections between OLC contigs and a GOV contig were weaker than connections between DBG contigs.

Using long-reads to evaluate the content of GIs: High quality, trimmed long reads from the WEC VirION dataset were mapped against the genomes of pelagiphage HTVC010P (NC_020481.1) and tig404 using minimap2 with the 'map-ont' preset and alignments were viewed in the Geneious v11.0.3 software. Reads with a minimum overlap of 100 bp with boundary regions of the GI were used in a tBLASTx query against the NR database via the NCBI website to evaluate function.

Assembly of tig404 with DBG and OLC: High quality, error-corrected short reads and raw long reads were mapped to the tig404 contig using bwa mem and minimap2, respectively. Mapped reads were extracted with samtools (https://github.com/samtools/) and used as inputs to hybrid DBG assembly with metaSPAdes[39] and OLC assembly with Canu[22]. Assembly graphs were then visualised with Bandage[49].

Acknowledgements The authors thank the crew of the Plymouth Marine Laboratory vessel 'Quest' for collection of seawater samples, as well as Dr Simon Roux and Dr Benjamin Bolduc for guidance and advice on bioinformatic analyses. Major support was provided by a fellowship to B.T. from the Bermuda Institute of Ocean Sciences as part of the BIOS-SCOPE program; the Royal Society and the Natural Environment Research Council (NERC) (NE/P008534/1 to B.T). JWD was funded by a NERC Great Western Four+ (GW4+) Doctoral Training Partnership PhD (NE/L002434/1).

Author contributions B.T. and J.W.D conceived and designed the experiments. N.S. and and L.C cultured mock viral community phages and recovered the DNA. JWD recovered DNA from the WEC viruses, produced and optimised VirION libraries from the mock viral community and the WEC viruses, and

sequenced them on the MinION. K.M. produced and sequenced Illumina libraries. B.T. conceived and ran bioinformatic analyses; microdiversity measures were calculated by A.G. JWD and BT wrote the manuscript and generated the figures, with contributions and editing from M.A, and editing by M.S.

Additional information Supplementary information is available online. Correspondence and requests for materials should be addressed to B.T. Sequencing data and assemblies are available at the European Nucleotide Archive under the project accession number PRJEB27181.

Competing interests The authors declare no competing financial interests

References

1. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **5,** 801–812 (2007).
2. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348,** (2015).
3. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537,** 689–693 (2016).
4. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8,** 15892 (2017).
5. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9,** e1003987 (2013).
6. Mizuno, C. M., Ghai, R., Saghaï, A., López-García, P. & Rodriguez-Valera, F. Genomes of Abundant and Widespread Viruses from the Deep Ocean. *MBio* **7,** (2016).
7. 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. PeerJ, 2017. doi:10.7717/peerj.3817
8. Mizuno, C. M., Ghai, R. & Rodriguez-Valera, F. Evidence for metaviromic islands in marine phages. *Front. Microbiol.* **5,** (2014).
9. Ashton, P. M. *et al.* MinION nanopore sequencing identifies the position and structure of a bacterial

antibiotic resistance island. *Nat. Biotechnol.* **33,** 296–300 (2015).

10. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36,** 338–345 (2018).
11. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6,** 100 (2017).
12. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28,** 127–181 (2004).
13. Forterre, P. The virocell concept and environmental microbiology. *ISME J.* **7,** 233–236 (2013).
14. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Curr. Opin. Microbiol.* **31,** 161–168 (2016).
15. Hurwitz, B. L., Hallam, S. J. & Sullivan, M. B. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol.* **14,** R123 (2013).
16. Breitbart, M. Marine Viruses: Truth or Dare. *Ann. Rev. Mar. Sci.* **4,** 425–448 (2012).
17. Brum, J. R. *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348,** 1261498 (2015).
18. Olson, N. D. *et al.* Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx098
19. Temperton, B. & Giovannoni, S. J. Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.* **15,** 605–612 (2012).
20. Zhao, Y. *et al.* Abundant SAR11 viruses in the ocean. *Nature* **494,** 357–360 (2013).
21. Laver, T. W. *et al.* Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Sci. Rep.* **6,** 21746 (2016).
22. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27,** 722–736 (2017).
23. Duhaime, M. B. *et al.* Comparative Omics and Trait Analyses of Marine *Pseudoalteromonas* Phages Advance the Phage OTU Concept. *Front. Microbiol.* **8,** 1241 (2017).
24. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial

genomic data. *PeerJ* **3,** e985 (2015).

25. Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5,** e3243 (2017).
26. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76,** 5269–5273 (1979).
27. Abbasifar, R. *et al.* Supersize me: *Cronobacter sakazakii* phage GAP32. *Virology* **460-461,** 138–146 (2014).
28. Šimoliūnas, E. *et al.* Genome of *Klebsiella sp.*-infecting bacteriophage vB_KleM_RaK2. *J. Virol.* **86,** 5406 (2012).
29. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5,** 69 (2017).
30. Luo, E., Aylward, F. O., Mende, D. R. & DeLong, E. F. Bacteriophage Distributions and Temporal Variability in the Ocean's Interior. *MBio* **8,** (2017).
31. Roux, S. *et al.* Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4,** e2777 (2016).
32. Noble, R. T. Enumeration of viruses. in *Methods in Microbiology* **30,** 43–51 (Academic Press, 2001).
33. Cunningham, B. R., Brum, J. R., Schwenck, S. M., Sullivan, M. B. & John, S. G. An inexpensive, accurate, and precise wet-mount method for enumerating aquatic viruses. *Appl. Environ. Microbiol.* **81,** 2995–3000 (2015).
34. 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).
35. Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15,** 1428–1440 (2013).
36. Deng, L. *et al.* Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513,** 242–245 (2014).
37. Bonilla, N. *et al.* Phage on tap-a quick and efficient protocol for the preparation of bacteriophage

laboratory stocks. *PeerJ* **4,** e2261 (2016).

38. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
39. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27,** 824–834 (2017).
40. Mikheenko, A., Saveliev, V. & Gurevich, A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32,** 1088–1090 (2016).
41. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).
42. Delcher, A. L., Salzberg, S. L. & Phillippy, A. M. Using MUMmer to identify similar regions in large sequence sets. *Curr. Protoc. Bioinformatics* **Chapter 10,** Unit 10.3 (2003).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
44. Tithi, S. S., Aylward, F. O., Jensen, R. V. & Zhang, L. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6,** e4227 (2018).
45. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).
46. Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.* **14,** e1002342 (2016).
47. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9,** 471–472 (2012).
48. Dabney, A., Storey, J. D. & Warnes, G. R. qvalue: Q-value estimation for false discovery rate control. *R package version* **1,** (2010).
49. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31,** 3350–3352 (2015).