Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands

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] and Ben Temperton[2]

[1] Plymouth Marine Laboratory, Plymouth, Devon, United Kingdom
[2] School of Biosciences, University of Exeter, Exeter, Devon, United Kingdom
[3] Department of Microbiology, Ohio State University, Columbus, OH, United States of America
[4] Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH, United States of America

ABSTRACT

Marine viruses impact global biogeochemical cycles via their influence on host community structure and function, yet our understanding of viral ecology is constrained by limitations in host culturing and a lack of reference genomes and 'universal' gene markers to facilitate community surveys. Short-read viral metagenomic studies have provided clues to viral function and first estimates of global viral gene abundance and distribution, but their assemblies are confounded by populations with high levels of strain evenness and nucleotide diversity (microdiversity), limiting assembly of some of the most abundant viruses on Earth. Such features also challenge assembly across genomic islands containing niche-defining genes that drive ecological speciation. These populations and features may be successfully captured by single-virus genomics and fosmid-based approaches, at least in abundant taxa, but at considerable cost and technical expertise. Here we established a low-cost, low-input, high throughput alternative sequencing and informatics workflow to improve viral metagenomic assemblies using short-read and long-read technology. The 'VirION' (Viral, long-read metagenomics via MinION sequencing) approach was first validated using mock communities where it was found to be as relatively quantitative as short-read methods and provided significant improvements in recovery of viral genomes. We then then applied VirION to the first metagenome from a natural viral community from the Western English Channel. In comparison to a short-read only approach, VirION: (i) increased number and completeness of assembled viral genomes; (ii) captured abundant, highly microdiverse virus populations, and (iii) captured more and longer genomic islands. Together, these findings suggest that VirION provides a high throughput and cost-effective alternative to fosmid and single-virus genomic approaches to more comprehensively explore viral communities in nature.

Subjects Bioinformatics, Genomics, Marine Biology, Microbiology, Virology
Keywords Viral Metagenomics, Virus, Virome, Metagenome, Assembly, Viral ecology, Long-read sequencing, Marine Microbiology

INTRODUCTION

The marine bacterial communities that regulate global carbon biogeochemical cycles are themselves structured by selective, phage-mediated lysis (*Weinbauer, 2004*; *Suttle, 2007*). Bacteria co-evolve with their phages and exchange genetic information, and phages even 'reprogram' hosts during infection so as to channel host metabolism towards phage replication (*Forterre, 2013*; *Hurwitz, Hallam & Sullivan, 2013*; *Hurwitz & U'Ren, 2016*). Over the last decade, the convergence of high throughput sequencing and the use of universal taxonomic marker genes for bacteria have revolutionised our understanding of microbial ecology (*Torsvik & Ovreaas, 2002*; *Treusch et al., 2009*; *Thompson et al., 2017*). Problematically, however, viral ecologists lack parallel approaches. First, PCR-amplified marker genes are limited to a narrow subset of the viral community, and require degeneracies and amplification conditions that undermine the quantitative nature of the data (*Sullivan, 2015*). Second, while short-read viral metagenomics studies to date have provided clues to viral function (e.g., virally encoded, host-derived central metabolism genes, known as Auxiliary Metabolic Genes: AMGs) (*Breitbart et al., 2007*; *Hurwitz, Hallam & Sullivan, 2013*), and first estimates of global viral gene abundance and distribution (*Brum et al., 2015*; *Roux et al., 2016a*), they suffer from technical limitations. This is because short-read assemblies are composites of populations 'features' (*Mizuno, Ghai & Rodriguez-Valera, 2014*), with successful assembly a function of coverage and branch resolution in assembly graphs (*Temperton & Giovannoni, 2012*; *Olson et al., 2017*). This limits our ability to assemble viral populations where multiple strains are abundant and microdiverse (*Roux et al., 2017*), as well as genomic regions of high diversity, such as genomic islands (GIs), which, in microbes, often contain niche-defining genes that drive ecological speciation (*Coleman et al., 2006*). In these latter regions, assembly is impeded by low coverage and/or repeat regions at the boundaries (*Mizuno, Ghai & Rodriguez-Valera, 2014*; *Ashton et al., 2015*).

These are not just technical limitations—emerging data suggests that these obstacles alter our understanding of viral roles on important taxa and global carbon biogeochemistry. For example, the globally dominant members of the chemoheterotrophic order Pelagibacterales comprise up to 25% of all bacterioplankton and are major contributors in the conversion of marine dissolved organic matter back to atmospheric CO_2 (*Giovannoni, 2017*). Their associated viruses dominate global oceans (*Zhao et al., 2013*; *Martinez-Hernandez et al., 2018*) and are likely to contribute significantly to carbon turnover in surface water by release of labile intracellular carbon during lysis (*Suttle, 2005*; *Suttle, 2007*). However, the genomes of viruses associated with Pelagibacterales contain numerous GIs and/or high microdiversity (*Zhao et al., 2013*; *Martinez-Hernandez et al., 2018*). Such features fragment genomes in short-read assemblies, which reduces representation following contig size-selection for downstream analyses (*Martinez-Hernandez et al., 2017*; *Roux et al., 2017*). Though single-virus genomics (*Martinez-Hernandez et al., 2017*) and fosmid-based approaches (*Mizuno et al., 2013*; *Mizuno et al., 2016*) can overcome such issues, these methods are technically challenging and costly to implement.

Alternatively, recent advances in long-read sequencing technology might be leveraged to better capture microdiverse viral populations and genomic islands. Such approaches can yield very long reads (>800 kbp) (*Jain et al., 2015*; *Jain et al., 2018*; *Loman, Quick & Simpson, 2015*), which would be long enough to capture complete genomes of double-stranded DNA bacteriophages ('phages') (10–617.5 kbp (*Mahmoudabadi & Phillips, 2018*)). At a minimum, such long reads could span genomic global- and local repeat regions, which tangle the De Bruijn Graph and fragment the assembly (*Koren & Phillippy, 2015*). Long reads may also overcome assembly challenges in regions of low coverage, to improve overall assembly of genomes from both cultured isolates (*Wick et al., 2017*) and metagenomics (*Frank et al., 2016*; *Driscoll et al., 2017*). It is also probable that long-read assemblies using overlap-layout-consensus would be less prone to microdiversity-associated fragmentation of genomes observed in De Bruijn Graph approaches (*Martinez-Hernandez et al., 2017*; *Roux et al., 2017*).

The challenge is that long-read technologies (both from PacBio and Oxford Nanopore) currently require large amounts of input DNA (micrograms; *Jain et al., 2018*, instead of nanograms commonly available from natural viral communities in seawater; *Hurwitz et al., 2013*). Furthermore, PacBio subreads and nanopore reads have high error rates (5–10%), with the former enriched in insertion errors and the latter enriched in insertion-deletion errors (*Weirather et al., 2017*). Indel errors shift the reading frame of the DNA sequence and confound gene-calling algorithms, artificially inflating the number of identified stop codons and producing shorter gene calls (*Warr & Watson, 2019*). This is a particular problem for viral metagenomics as the median length of genes in dsDNA phages is approximately half that of their bacterial hosts (408 bp vs 801 bp, respectively) (*Brocchieri & Karlin, 2005*; *Mahmoudabadi & Phillips, 2018*), and the vast majority of viral genes in both dsDNA viral isolates and viral metagenomes (>50% and up to 93%, respectively) have no known function (*Hurwitz & Sullivan, 2013*; *Mahmoudabadi & Phillips, 2018*), making it difficult to evaluate the quality of gene calls from metagenomic assemblies.

Here, we adapted a Long-Read Linker-Amplified Shotgun Library (LASL) approach for quantitative viral metagenomics (*Duhaime et al., 2012*) to obtain sufficient quantities of high-molecular weight DNA from nanograms of viral community dsDNA for sequencing using the MinION sequencer from Oxford Nanopore Technology. We then established a bioinformatic workflow to combine such long-read data with complimentary short-read sequencing data to maximise the advantages and minimise the weaknesses of both sequencing technologies. Following validation on mock viral communities, we applied our new approach to the first marine viral metagenome from the Western English Channel (WEC). Here, we present the first use of long-read sequencing technology for viral metagenomics and show that this novel approach provides significant benefits when combined with short-read metagenomics.

MATERIALS & METHODS

Construction of the mock viral community

A mock viral community comprised of six isolated and sequenced marine Caudovirales with genome sizes ranging from 38.5–129.4 kbp was produced as described previously (*Roux et*

_al., 2016b_). 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 0.2 µm 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) (_Noble, 2001_; _Cunningham et al., 2015_). 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).

Construction of the Western English Channel viral metagenome

A total of 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 h of collection) via sequential filtration through glass fibre (GF/D: pore size 2.7 µm) and polyethersulfone (pore size 0.22 µm) filters in a 142 mm polycarbonate rig, with peristaltic pump. Precipitation of viruses from filtrate (denoted as the viral fraction) and primary concentration of virus particles was conducted by iron chloride flocculation and collection on 1.0 µm polycarbonate filters (_John et al., 2011_); filters were 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) (_Hurwitz et al., 2013_) that had been pre-treated with 1% bovine serum albumin buffer to minimise capsid-filter adhesion (_Deng et al., 2014_) 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 (_Bonilla et al., 2016_) and purified with DNase I (100 U/mL; 2 h 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 (_Hurwitz et al., 2013_). Viral DNA was extracted from concentrated and purified viral particles using the Wizard® DNA Clean-up System (Promega A7280), removing PCR inhibitors (_John et al., 2011_).

Library preparation, amplification and sequencing

For short-read sequencing, Illumina libraries were generated from 1 ng of either mock viral community DNA (Table S1), or 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).

For VirION libraries (Fig. 1), 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). End-repair of DNA fragments, amplification of DNA with
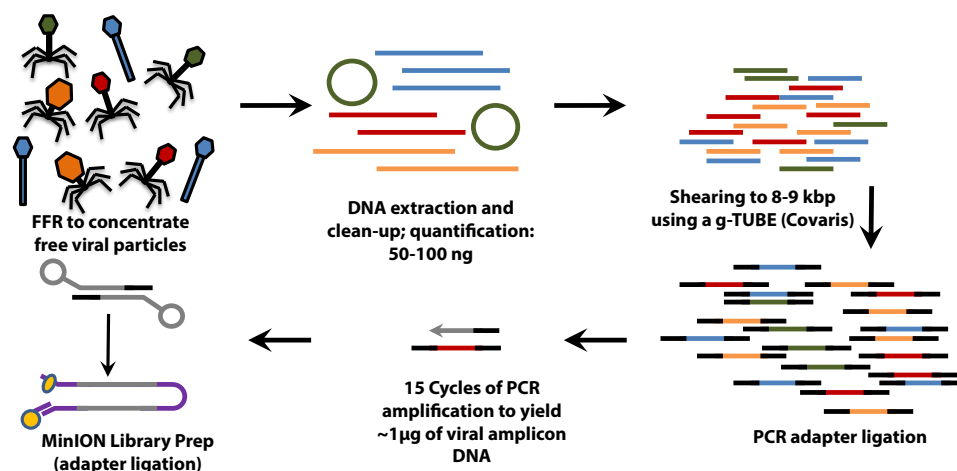
Figure 1 **Workflow for preparation of free-viral fraction DNA for MinION sequencing.** The long-read viral metagenomic method (VirION) developed includes $FeCl_3$ flocculation and resuspension (FFR), shearing of extracted viral DNA (to 8–9 kbp), random linker amplification (Linker Amplified Shotgun Library: LASL), MinION library preparation, and nanopore (Oxford Nanopore Technologies; ONT) sequencing.

Full-size DOI: 10.7717/peerj.6800/fig-1

PCR-adapter ligation (i.e., Linker Amplified Shotgun Library: LASL preparation), and preparation of MinION-compatible libraries were performed following the manufacturer's protocols for ''2D Low input genomic DNA with PCR'' using the 'Ligation Sequencing kit 2D' (ONT SQK-LSK208). PCR reaction conditions were modified with reference to NEBNext High-Fidelity 2X PCR Master Mix (NEB M0541S) manufacturer's instructions in order to maximise DNA yield, whilst minimising production of chimeric sequences, as follows: 3 m at 95 °C (initial denaturation), 15 cycles of: 15 s at 95 °C (denaturation), 15 s 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. ∼1.5 μg of end-repaired, amplified DNA was carried forward for sequencing adapter ligation followed by purification of adapted DNA using MyOne C1 Streptavidin beads (Thermo Fisher Scientific Inc. 65001). The prepared long read library was sequenced on a single MinION Mk 1B flow cell with R9.4 pore chemistry for 48 h (Note—to remain up to date with changing ONT chemistry, a 1D ligation version of this protocol has also been tested and is available on protocols.io (https://www.protocols.io/view/virion-long-read-low-input-viral-metagenomic-sequ-p8fdrtn). Quality control of short and long read libraries was performed as described in Supplemental Information. Hiqh quality sequence data were used to generate short-read De Bruijn Graph assemblies (using metaSPAdes v. 3.11; *Nurk et al., 2017*), hybrid long-read scaffolded De Bruijn Graph assemblies (using metaSPAdes, with—nanopore parameter), and long-read overlap-layout consensus assemblies (with Canu; *Koren et al., 2017*) following optimisation for metagenomic data (see Supplemental Information). Rates of chimerism in both VirION reads (formed during PCR amplification) and assemblies (formed by mis-assembly) were evaluated by aligning reads and contigs, respectively, from the mock viral community to their associated genomes (Table S1).

Maximizing the benefits of long read and short read assemblies

We developed a bioinformatic pipeline to maximise the benefits of VirION reads for viral metagenomics (Fig. 2). Briefly, long-read assembly contigs of VirION reads were 'polished' with matching short-read sequences to remove sequencing error via consensus base-calling (using Pilon; *Walker et al., 2014*, v1.22). In order to capture the longest assemblies available from the short-read data, scaffolds from short-read and hybrid assemblies were combined and dereplicated using a cut-off of 95% average nucleotide identity over 80% of the length (via MUMmer v3.23; *Delcher, Salzberg & Phillippy, 2003*) to cluster highly similar contigs into viral populations (*Roux et al., 2016a*). The longest representatives of each population were carried forward for analysis. Population representatives >10 kbp were pooled with polished long-read assembly contigs >10 kbp and evaluated with VirSorter (*Roux et al., 2015*) (in virome decontamination mode) to identify putative viral contigs. Reads classified as category 3 (deemed unusual, but not necessarily viral; *Roux et al., 2015*) were excluded from downstream analyses. Circular contigs (i.e., where the contig has matching ends) were identified by VirSorter and used as a proxy for successful assembly of a complete genome. Matching short-read data was then mapped against the representative viral population sequences (using bowtie2; *Langmead & Salzberg, 2012*) for use in evaluating (1) relative abundances of contigs; (2) whether long read assembly captured more microdiverse genomes; and (3) recovery of genomic islands and their predicted functional composition at the population level (see Supplemental Information).

To direct future sampling efforts in environmental samples, we evaluated the short-read sequencing depth at which inclusion of long reads in hybrid assemblies offered no advantage in genome recovery. High-quality short read sequences were randomly subsampled in triplicate to seven discrete depths representing 10% and 70% of the full dataset (using seqtk https://github.com/lh3/seqtk). Subsampled reads were then assembled with and without scaffolding support from VirION reads (with metaSPAdes; *Nurk et al., 2017*). Scaffolds >10 kbp in replicated assemblies were classified as viral using VirSorter (*Roux et al., 2015*) in virome decontamination mode and the number of scaffolds classified as viral 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 triplicate replicates at each sequencing depth.

Validation of error correction of long reads in viral metagenomic data

We evaluated whether it was possible to use short-read data to correct base-calling errors in long-read environmental metagenomic data in a similar way to that used for genomes of bacteria and eukaryotes from axenic samples (*Walker et al., 2014*). Western English Channel short-read data were sub-sampled to different sequencing depths, in triplicate; sub-samples were then mapped against the long-read assemblies of VirION reads. For error-correction with Pilon (*Walker et al., 2014*) and median coverage, the total number of fixed deletions and fixed insertions at each coverage depth were calculated. We then evaluated whether error-correction could be used to reduce the impact of frameshift errors on predicted gene length. Predicted coding sequences were identified using MetaGeneAnnotator (*Noguchi, Taniguchi & Itoh, 2008*) on the following: (1) uncorrected VirION reads; (2) long-read
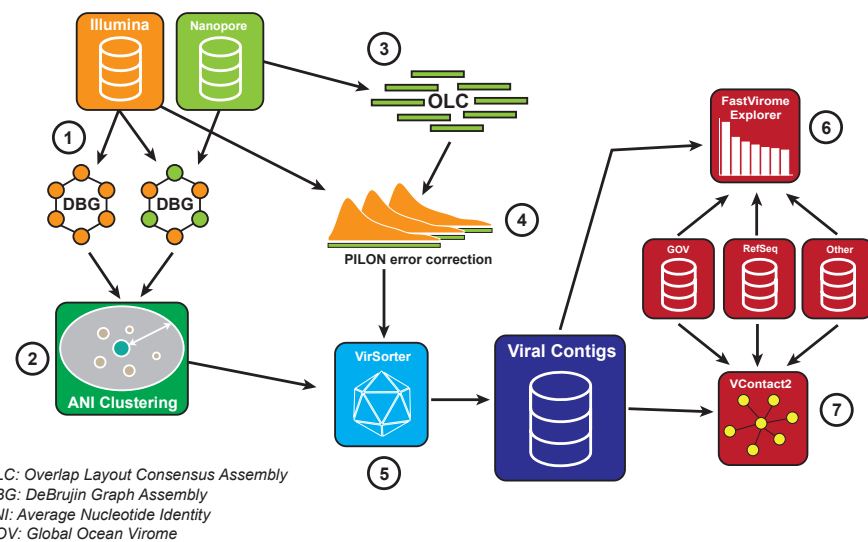
OLC: Overlap Layout Consensus Assembly
DBG: DeBrujin Graph Assembly
ANI: Average Nucleotide Identity
GOV: Global Ocean Virome

Figure 2 **Bioinformatics pipeline for VirION reads and complementary short-read sequencing for viral metagenomes.** The VirION bioinformatic pipeline to combine for short-read (Illumina) and long-read (MinION) sequencing to maximise the advantages of both sequencing platforms. Viral metagenomic short-read data and VirION reads from the Western English Channel were processed for identification of putative viral genomes as follows: (1) Short-read contigs and contigs scaffolded with VirION reads were generated via De Bruijn Graph Assembly using metaSPAdes (*Nurk et al., 2017*), and (2) de-replicated via average nucleotide identity of 95% similarity across 80% length. Separately, (3) long, error-prone VirION reads were assembled via overlap layout consensus Assembly using Canu (*Koren et al., 2017*) and (4) error-corrected via alignment of Illumina reads and consensus base calling with Pilon (*Walker et al., 2014*). (5) Putative viral genomes were identified using VirSorter (*Roux et al., 2015*). (6) Relative and global abundances of the Western English Channel viral contigs were calculated via competitive recruitment of short read data with FastVirome Explorer (*Tithi et al., 2018*), and lastly, (7) viral clusters based on shared proteins were produced from Western English Channel viral contigs clustered with contigs from the Global Ocean Virome (*Roux et al., 2016a*) and NCBI's RefSeq database (v.8.4 among others—see Table S2) using vConTACT2 (*Bolduc et al., 2017*).

Full-size ◨ DOI: 10.7717/peerj.6800/fig-2

assemblies of VirION reads; (3) long-read assemblies of VirION reads polished with the full short-read dataset; (4) contigs from scaffolded short-read assemblies; (5) contigs from the hybrid assembly. Distributions of the lengths of predicted coding sequences were compared against those in the genomes of *Caudovirales* from the NCBI RefSeq database (v.8.4), predicted proteins from the Global Ocean Virome (GOV; *Roux et al., 2016a*), and the single-amplified viruses in *Martinez-Hernandez et al. (2017)*. Effect size of different assembly types on genomic island length and density and associated 95% confidence intervals (CI) were calculated from bootstrapped medians (*Cumming, 2014*). For each bootstrap, 1000 predicted proteins were randomly subsampled from each dataset and their median length was calculated.

Analysis of Tig404—a contig closely related to Pelagiphage HTVC010P

Phage contigs closely related to Pelagiphage HTVC010P were identified via clustering of viruses at the ICTV-accepted level of genera by shared gene content (vContact2; *Bolduc et al., 2017*). Within this viral cluster, contig *tig404*, from the polished long-read assembly

of VirION reads, was identified as a circular viral contig by VirSorter (*Roux et al., 2015*). Whole genome alignment was performed with MUmmer (*Delcher, Salzberg & Phillippy, 2003*) to calculate average nucleotide identity to HTVC010P. Contigs from short-read only and hybrid assemblies that shared 95% nucleotide identity over 80% of their length to tig404 were identified and mapped back to their respective loci with MUmmer (*Delcher, Salzberg & Phillippy, 2003*). Genomic islands and nucleotide diversity of tig404 were calculated as described previously. To evaluate the contents of a 5.3 kb genomic island, unpolished VirION reads were mapped back against the tig404 genome and those which mapped to at least 100 bp on the borders of the genomic island were extracted. Mapped reads extending at least 1 kb into the genomic island were used as a query in a tBLASTx best-BLAST (*Camacho et al., 2009*) search against the NCBI NR database to annotate the reads whilst minimising the adverse impact of sequencing error within the uncorrected reads.

Estimating relative abundance and viral clusters of WEC viruses in viral metagenomes

FastViromeExplorer (*Tithi et al., 2018*) v.1.1 was used to quantify the relative abundances of WEC viral contigs. FastViromeExplorer is built upon the Kallisto (*Bray et al., 2016*) 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 Global Ocean Virome (*Roux et al., 2016a*) and from our Western English Channel sample were randomly subsampled to 10 million reads using seqtk (https://github.com/lh3/seqtk) to standardise per-sample sequencing effort, with the number of reads selected to balance detection of lower abundance viral populations with maximising the number of samples that could be included in the survey. Subsampled reads for each sample were recruited against a Kallisto index comprising (1) The viral genomes >10 kbp identified in this study; (2) A selection of phage genomes >10 kbp from key metagenomic studies (*Roux et al., 2016a*; *Martinez-Hernandez et al., 2017*; *Luo et al., 2017*); (3) Cultured viruses from the NCBI RefSeq viral database (v8.4) (Table S2). Contigs >10 kbp were selected to maximise accuracy of VirSorter to correctly identify viral contigs (*Roux et al., 2015*). For inclusion in downstream abundance analyses, contigs with less than 40% coverage as calculated by FastViromeExplorer were classified as having zero abundance to avoid over-representation of partial matches (*Tithi et al., 2018*). The top 100 most abundant contigs from each sample were also selected for downstream analyses. All phage genomes >10 kbp (including those from RefSeq) were processed using VirSorter (v.1.03) on the CyVerse cyberinfrastructure (*Merchant et al., 2016*) to standardise gene-calling prior to clustering of viruses into ICTV-recognised genera by shared gene content using vContact2 (*Bolduc et al., 2017*). In the final stage of clustering, vContact2 uses ClusterONE (*Nepusz, Yu & Paccanaro, 2012*) 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 (*Dabney, Storey & Warnes, 2015*) to account for multiple testing and a *q*-value cutoff of <0.05 was used to identify statistically significant clusters.

RESULTS & DISCUSSION

Here, we present the first use of long-read sequencing technology for viral metagenomics and show that this novel approach provides significant benefits when combined with short-read metagenomics. Our bioinformatics pipeline overcame the high sequencing error associated with long-read technology and the addition of long reads enabled capture of complete viral genomes which were globally ubiquitous, and not represented by short-read only assemblies. Long-read assemblies also significantly improved the capture of viral genomic islands, demonstrating that this advance will facilitate better understanding of niche-differentiation and ecological speciation of viruses in environmental samples.

Assembly of VirION reads successfully captured mock viral community genomes and retained relative abundance information

VirION sequencing of the mock viral community produced 359,338 high quality (Q>10) long reads (median length: 4,099 bp; max length 18,644 bp). 95% of the reads (341,718) mapped back to the genomes of the mock viral community. Considering viral DNA was sheared to 6–8 kbp fragments, the length of amplicons following LASL were shorter than expected, presumably due to preferential PCR amplification of shorter fragments (*Shagin et al., 1999*) (Fig. S1A) and preferential diffusion (and thus sequencing) of shorter reads within the flowcell microfluidics (Fig. S1B). Only 0.95% of LASL amplified reads were classified as chimeric (mapping to more than one location of the same or different genomes of the mock viral community), suggesting 15 rounds of PCR was sufficiently low to minimise production of chimeric artifacts, supporting previous findings (*Laver et al., 2016*). Several methods have been developed for sequencing dsDNA viral metagenomes without skewing relative abundance information important for comparative ecology, including an LASL approach optimised for 454 sequencing (*Duhaime et al., 2012*; *Hurwitz et al., 2013*) and Nextera sequencing (*Roux et al., 2016b*). Median per-genome coverages of VirION reads and short-read Nextera datasets (5.6 M 2 × 300 bp paired-end) from the mock viral community were strongly correlated ($R^2 = 0.975$, $p < 0.001$, Fig. 3A), indicating that the LASL approach used here for multi-kilobasepair fragments retained relative abundance information observed in previous LASL approaches.

Long-read and short-read assemblies of the mock viral community captured >99.7% of the six mock viral community genomes (Table S1). Neither the short-read only assembly, hybrid assembly nor long-read assemblies were able to capture all six genomes in six complete contigs. Long-read methods gave the most contiguous assemblies, capturing the six genomes across 14 contigs. In comparison, short-read only assemblies recovered the genomes across 26 contigs, whereas hybrid assembly reduced the number of contigs to 21. As expected, we identified >250 times more indel errors in long-read only assemblies than in the short-read assemblies scaffolded with long reads (average of 474 vs <2 indels per 100 kbp, respectively). Polishing of long-read only assemblies with short read data reduced the indel error rate to 22.78 per 100 kbp, indicating this was a successful strategy for reducing indel error of long-read assemblies in metagenomic samples, but was not able to remove such errors completely. There was no evidence of chimerism in any of the assemblies, indicating that Canu's *in silico* correction of chimeras (*Koren et al., 2017*)

successfully removed the low number of chimeric sequences observed in the VirION reads during assembly.

Combining VirION reads with short read data improves viral metagenomic assembly in an environmental virome

Long read sequencing of an environmental virome from the Western English Channel produced 108,718 high quality VirION reads (median length: 3,625 bp; max length: 17,019 bp, total yield of 0.39 Gbp). It is worth noting that recent developments of MinION technology have improved flowcell yields to >10 Gbp (*pers comms*). Therefore, our analyses here represent low coverage of the viral community with long read data compared to currently available (and fast-improving) technology.

Scaffolding short-read assemblies using VirION reads provided a small, but significant increase in the number of putative viral genomes recovered (between 1.1 to 1.5-fold increase, Student t-test, $p < 0.05$) than short-read only assemblies up to a short-read sequencing depth of ∼12 Gbp (Fig. 3B). Above this depth, there was no significant difference between short-read assemblies with and without scaffolding, suggesting assembly of short-read data was capturing most of the viral community above this sequencing depth. For comparison, the median sequencing depth of 137 Illumina sequenced viral metagenomes from the Global Ocean Virome survey (study PRJEB27181 in the European Nucleotide Archive) was 8.67 Gbp (IQR = 5.22 Gbp), with 110 out of 137 samples sequenced to a depth of <12 Gbp. Inclusion of VirION reads in hybrid assemblies significantly increased the number of 'complete' (i.e., circular contigs) viral genomes recovered once short-read sequencing depth increased above 12 Gbp (1.5 to 2.0-fold, Student t-test, $p < 0.05$) (Fig. 3B). Details of differences in means and p-values at each depth are available in Tables S3 and S4. When the full (30.8 Gbp) short-read dataset was used, the inclusion of long reads for scaffolding De Bruijn Graph assemblies increased the median length of recovered viral genomes by an average of 1.8 kbp compared to short-read only assemblies (Mann–Whitney U test, $n_1 = 1,400$, $n_2 = 879$, p-value < 0.001). With an estimated mean gene density of 1.4 genes per kb in phage dsDNA genomes (*Mahmoudabadi & Phillips, 2018*), this increased length represents an extra 2.5 genes per contig.

Polishing of long-read assemblies of WEC VirION reads using complementary short-read data removed a maximum of 172,854 insertion errors and 12,674 deletion errors (Fig. S2). Error correction reached an asymptote at ∼9 Gbp of short-read sequencing data, with a median coverage of ∼70. As expected, the errors associated with long-read sequencing adversely affected the lengths of protein predictions (Fig. S3), in accordance with previous findings (*Warr & Watson, 2019*). Proteins predicted from uncorrected VirION reads (median length of 72 aa, 70–74 aa 95% CI) were shorter (median difference = 61 aa, 69–53 aa 95% CI) than those from RefSeq Caudovirales genomes (median length of 133 aa, 126–141 aa 95% CI), and much shorter (median difference = 88 aa, 83–95 95% CI) than those from the GOV dataset (median length of 160 aa, 149–173 95% CI). Assembly of long reads with Canu includes a consensus-based error-correction step (*Koren et al., 2017*), which increased median predicted protein lengths to 87 aa (median difference of 15 aa, 14–15 95%CI) compared to raw VirION reads. Polishing of long-read assemblies of
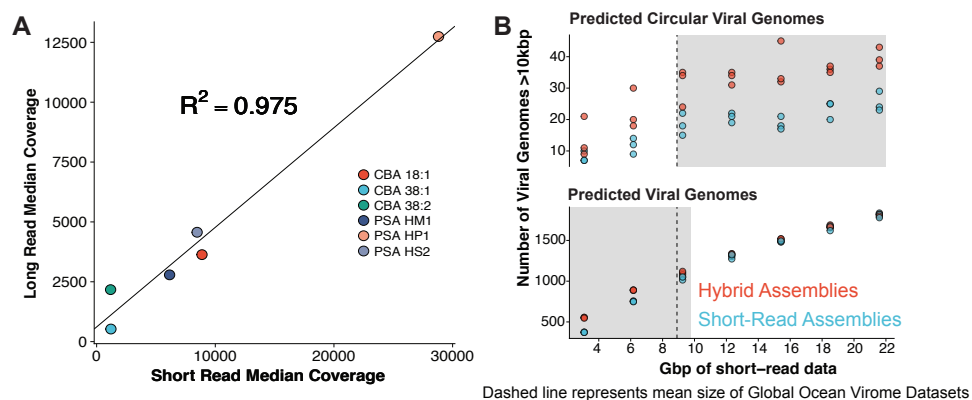
Warwick-Dugdale et al. (2019), *PeerJ*, DOI 10.7717/peerj.6800

10/28

Figure 3 **Comparative performances of short-read and long-read data for the identification of marine viral genomes.** (A) Relative abundances of genome-mapped VirION reads and short-reads 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 strongly correlated using both approaches, showing amplification of sheared viral DNA for VirION sequencing was as quantitative as short read approaches 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 random subsamples of short read data from the Western English Channel viral metagenome: 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).

Full-size 🖼 DOI: 10.7717/peerj.6800/fig-3

VirION reads with short read data was highly effective in restoring the length of predicted proteins (median length 127 aa, 120–135 aa 95%CI) to lengths similar to those observed in RefSeq Caudovirales (median length = 133 aa, 126–141 aa 95%CI). Proteins from polished reads had a median difference of −6 aa (−18–6 95%CI) compared to RefSeq Caudovirales proteins. This suggests that not all frameshift errors were corrected in the long-read assemblies, corroborated by evidence of increased indel errors observed in long-read assemblies of mock viral community data compared to short-read assemblies.

Interestingly, predicted protein lengths from the GOV dataset (*Roux et al., 2016a*) (median length = 160 aa), short-read only assembly of the WEC virome (median length = 157 aa); hybrid assembly of the WEC virome (median length = 160 aa) and data from single-amplified viral genomes (*Martinez-Hernandez et al., 2017*) (median length = 152 aa) were all of similar length and 19 to 27 aa longer compared to those from RefSeq Caudovirales genomes, and 25 to 33 aa longer than those from WEC polished long-read assemblies. In comparison, median predicted protein length in 899 dsDNA phages was previously estimated at 136 aa (*Mahmoudabadi & Phillips, 2018*)—similar to those found in our polished long-read assemblies from VirION reads. Thus, either both the RefSeq Caudovirales dataset and that of Mahmoudabadi and Phillips are under-representing longer viral predicted proteins found in marine viral metagenomes, or predicted protein lengths in viral genomes from metagenomic data are longer than those observed in cultured representatives. Whether this difference is biological or an artifact of metagenomic assembly and gene calling is an interesting area for further investigation.

Assembly and mapping of VirION reads captures more information about potential niche-defining genomic islands than short-read only or hybrid assemblies: In marine bacteria, genomic islands have been identified as playing an important role in niche specialisation that drives ecological speciation (*Coleman et al., 2006*). Genomic islands have also been found to be a common feature of viral genomes and are typically enriched in functions associated with host recognition (*Mizuno, Ghai & Rodriguez-Valera, 2014*). At all nucleotide identity cut-offs tested, genomic islands captured on long-read assemblies were between 145 bp (112–184 bp 95%CI) and 225 bp (189–259 bp, 95% CI) longer than those captured on short-read only or hybrid assemblies. (Fig. 4A, Fig. S4A). There were no significant differences between the lengths of genomic islands captured on short-read only or hybrid assemblies. The largest genomic islands in each assembly type were 2.47 kbp, 5.75 kbp and 5.65 kbp in short-read only assemblies, hybrid assemblies and long-read assemblies, respectively. In comparison, the largest genomic islands identified in fosmid-based viral metagenomes were ~4.6 kbp (*Mizuno, Ghai & Rodriguez-Valera, 2014*), suggesting that both hybrid and long-read approaches capture similar length genomic islands as previous fosmid-based methods. Similarly, the density of GIs was significantly greater in long-read assemblies (at between 40 bp (20–60 bp, 95%CI) and 100 bp (80–110 bp, 95%CI) of GI per kbp of genome) compared to short-read or hybrid assemblies (Fig. 4B, Fig. S4B). Again, there was no significant difference between short-read only and hybrid assemblies. At a nucleotide identity cut-off of 98% for read mapping, the length of GIs in long-read assemblies were longer than those at 92% and 95%, (59 bp (18–106 bp, 95%CI) and 61 bp (13–105 bp, 95%CI) respectively), indicating that residual error in the polished reads may be contributing to a slight increase in predicted GI length and density at high nucleotide identity. However, these effect sizes are much smaller than those observed between long-read assemblies and short and hybrid assemblies across all identity cut-offs, suggesting that long reads do indeed improve the capture of genomic islands.

Previous work to identify viral genomic islands by recruiting short-reads back against assembled contigs from fosmid libraries showed that nucleotide diversity within genomic islands was associated with a constant-diversity model, with under-recruiting islands containing proteins associated with host recognition and penetration, phage structure and DNA packaging structural proteins (*Mizuno, Ghai & Rodriguez-Valera, 2014*). Here, we were able to gain further insight into viral genomic islands by investigating whether diversity also occurred at the functional level. VirION reads that spanned the full width of GIs were identified and their gene content was predicted at the nucleotide level. Reads spanning the same GI were compared to see if different proteins were encoded on different template strands prior to amplification.

In total, 137 genomic islands on 84 viral contigs had at least 10 VirION reads spanning their full length. The 3,072 reads spanning these islands encoded 6,445 predicted proteins, of which 4,599 could be aligned to a protein within the NR database. Just 711 (15%) of aligned predicted proteins returned a hit with known function, indicating that genomic islands are an important source of 'genetic dark matter' (i.e., sequence of unknown function) in viral metagenomes (*Krishnamurthy & Wang, 2017*). In total 66 genomic islands contained genes with an assigned function. These islands captured a range of functional proteins
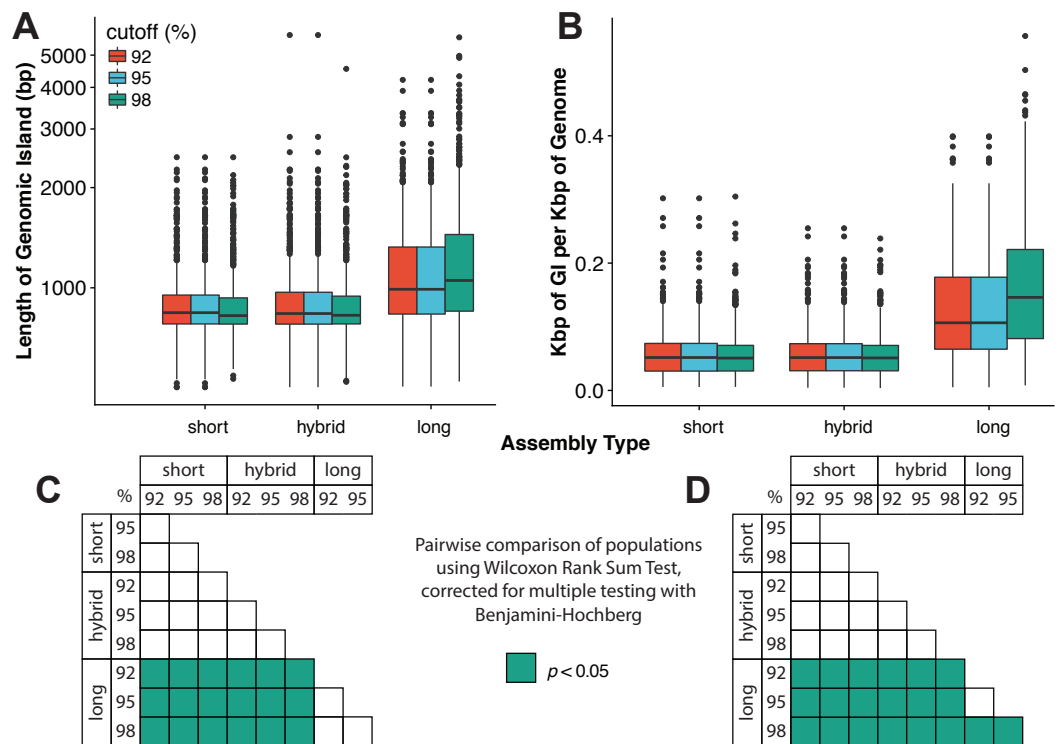
Figure 4 Long-read assemblies capture longer genomic islands than short-read methods. Comparison of the (A) length of genomic islands (GI) and (B) normalised length of GI per kb of genome per contig captured on long read assemblies of VirION reads compared to short- read only and hybrid assemblies of viral contigs from the Western English Channel. Genomic islands were identified by mapping reads back against contig across a range of nucleotide percentage identities (92, 95, 98%) to account for residual error remaining in polished long- read assemblies. (C) and (D) represent pairwise significance calculated using a Wilcoxon Rank Sum Test, with p-values adjusted (Benjamini-Hochberg) for multiple testing, for (A) and (B), respectively. Effect sizes and 95% confidence intervals can be found in Fig. S4.

Full-size 🖼 DOI: 10.7717/peerj.6800/fig-4

including those associated with nucleotide biosynthesis; DNA methylation; redirection of host machinery; structural proteins and associated chaperonins; endo and exonucleases and integrases (Table S5). 35 of the genomic islands contained structural proteins (capsid, tail proteins, co-chaperonin GroES, YapH); proteins associated with membrane recognition (carbohydrate-binding module, lectin-binding proteins) or proteins associated with reconfiguring host metabolic machinery for viral synthesis or defence suppression (RNA polymerase sigma factor, methyltransferases, tRNA synthetases, anti-restriction protein), supporting the hypothesis that viral genomic islands are a hotspot for Constant-Diversity evolutionary dynamics (*Mizuno, Ghai & Rodriguez-Valera, 2014*). Seven out of eight genomic islands containing a thymidylate synthase (an enzyme involved in pyrimidine metabolism) also encoded partial hits to ribonucleotide reductase (involved in both purine and pyrimidine metabolism). Ribonucleotide reductase has previously been identified as the nucleotide metabolism gene most frequently interrupted by self-splicing introns (*Dwivedi et al., 2013*). Similarly, thymidylate synthase has been found to contain self-splicing, group I introns in phage genomes (*Chu et al., 1984*; *Bechhofer, Hue & Shub, 1994*), potentially

identifying intron splicing activity as a source of regulatory/functional variability and/or as a mechanism to promote the movement of genetic material within the viral genomic islands in our data. Functional, putatively niche-defining metabolic genes were also identified in the genomic islands, including an ultraviolet light damage repair gene *uvsE* and genes associated with photosystem II *(psbA)*. Twenty-five out of the 66 genomic islands showed evidence of alternative gene arrangements across their spanning reads, suggesting the content of genomic islands can vary within viral populations at the structural, functional and nucleotide level.

Assembly of VirION reads capture important, microdiverse populations previously missed by short-read data

It has been hypothesised that genomes assembled from short-read metagenomes may be biased away from microdiverse populations (*Martinez-Hernandez et al., 2017*; *Roux et al., 2017*). We reasoned that overlap layout consensus assembly of long reads, followed by error correction might better capture genomes with high levels of microdiversity by avoiding the unresolvable branches of De Bruijn Graph assemblies. We evaluated genome-level nucleotide diversity (π) (*Nei & Li, 1979*) of both short-read assemblies and polished long-read assemblies from the Western English Channel virome. Median levels of π were significantly (3-fold) higher in polished long-read contigs than those derived from De Bruijn Graph assemblies (two-sided Mann–Whitney U test: $W = 105,830$, $n_1 = 758$, $n_2 = 206$, $p = 4.81 \times 10^{-15}$; Fig. S5), consistent with the hypothesis that long-read assembly of VirION reads captured genomes previously lost due to failure to resolve assembly graphs as a consequence of microdiversity.

Tig404—an example of how VirION reads improve viral metagenomics

The benefit of using VirION reads for viral metagenomics is exemplified by a polished contig from long-read assemblies that showed high nucleotide similarity and shared gene content to the globally abundant pelagiphage HTVC010P (*Zhao et al., 2013*). This ecologically important virus and its closely associated phages contain numerous genomic islands that comprise ∼10% of their genome and a shared 5.3 kbp genomic island containing a putative ribonuclease, bounded by tail fibre proteins (*Mizuno, Ghai & Rodriguez-Valera, 2014*). It has also been predicted to possess high microdiversity that challenges assembly from short-read data, leading to fragmentation and thus under-representation in short-read viral metagenomes, but is successfully captured using fosmid approaches and single-virus genomics (*Mizuno et al., 2013*; *Martinez-Hernandez et al., 2017*). Clustering of viral contigs from the WEC by shared-gene content using vContact2 (*Bolduc et al., 2017*) identified a virus called 'tig404' from long-read assembly of VirION reads that was 89% identical at the nucleotide level to HTVC010P. We mapped contigs from short-read only and hybrid assemblies against this genome at 95% nucleotide identity over 80% of the length to evaluate the success of short-read and hybrid assembly methods at capturing this genome, and identified its genomic islands as described above (Fig. 5). Both short-read only and hybrid assemblies were highly fragmented across the genome. Analysis of median nucleotide diversity of tig404 was extremely high (Fig. S5) and provided supporting evidence that
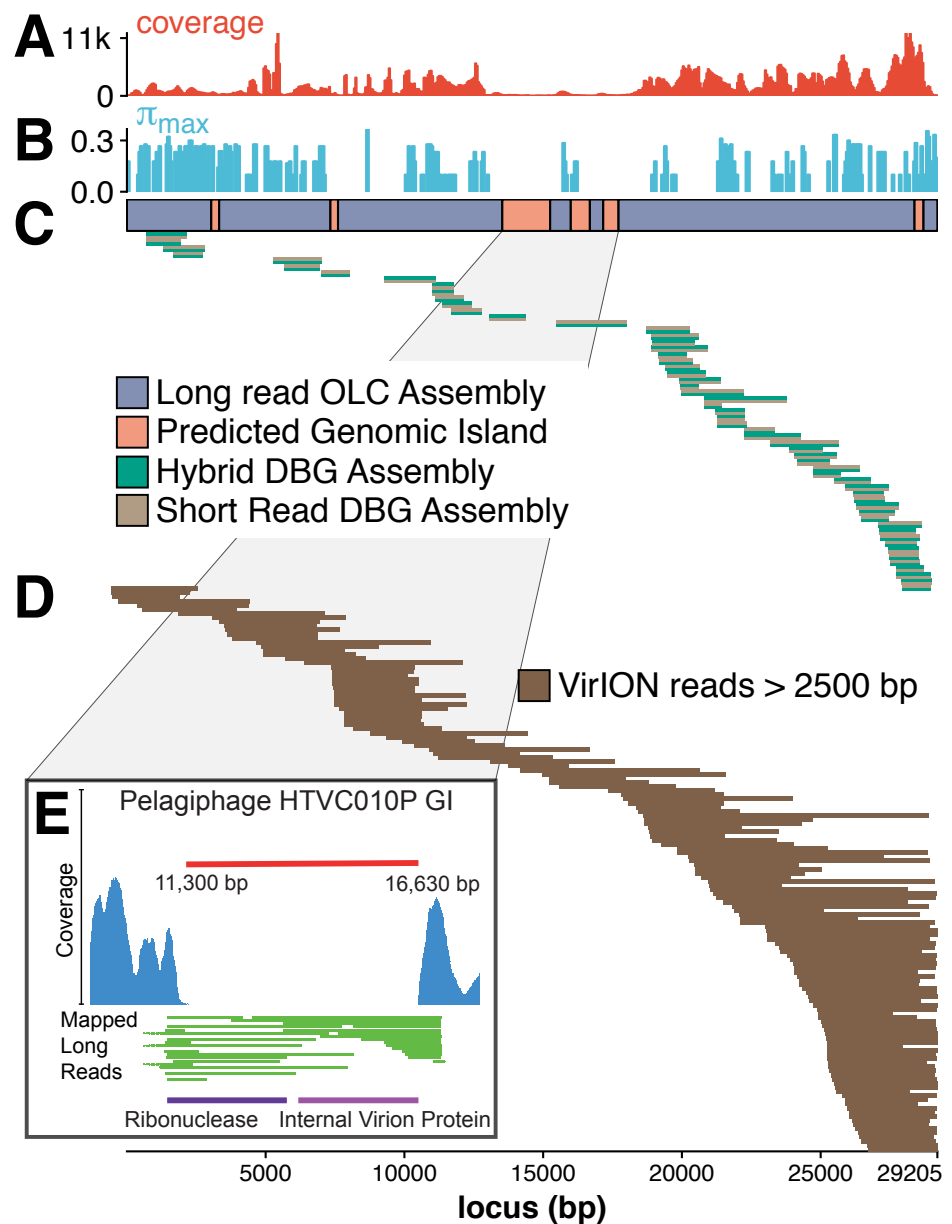
Figure 5 **Long-read sequencing resolves microdiversity and assembly issues across genomic islands in ecologically important viral taxa.** De Bruijn Graph (DBG) assembly of short reads, even with VirION reads for scaffolding failed to assemble the genome of tig404, a virus closely related to the globally abundant pelagiphage HTVC010P. Only long-read assembly of VirION reads, followed by error correction with short read data was able to capture the complete genome on a single 29.2 kbp contig. A 200 bp sliding window analysis was used to calculate median coverage (A) of the assembly and (B) maximum nucleotide diversity (π), revealing six genomic islands (GIs) (C) and high levels of nucleotide diversity. The impact of this on short-read (light brown) only and hybrid assembly (green) can be seen in (C), where the assemblies aligned to the long-read assembly are highly fragmented. Conversely, long VirION reads (dark brown) were capable of spanning these regions across the whole genome and thus enabling assembly (D). One genomic island on tig404 was conserved with that of HTVC010P (E). Thus, we were able to identify the genomic content of this island at the population level by mapping VirION reads to HTVC010P and identifying those that spanned the genomic island. Encoded function was then predicted using tBLASTx to overcome high sequencing error in uncorrected VirION reads.

Full-size 🖾 DOI: 10.7717/peerj.6800/fig-5

fragmentation may be a result of high microdiversity in this phage. In contrast, VirION reads successfully overlapped across the genome and enabled recovery of the genome through long-read assembly. Comparison of the genome of tig404 with that of HTVC010P identified a shared genomic island containing a putative ribonuclease protein and bounded by a tail fibre protein (Fig. 5), similar to those observed in closely related taxa from fosmid libraries (*Mizuno, Ghai & Rodriguez-Valera, 2014*).

In addition, we were able to exploit an additional benefit of long reads and use unpolished VirION reads to explore the contents of the shared genomic island across the tig404 population within the WEC virome. As each read is derived from a single DNA strand (excluding the low abundance of chimeric reads), variance in the content of the genomic island within a population would be captured on reads that align to the ends, or across, the genomic island. In total, 31 VirION reads extended from the boundaries into the genomic island (Fig. 5). Of these, 17 had sufficient overlap to use for identifying functional genes. Those at the 5′ end of the genomic island all contained a putative ribonuclease, whilst those at the 3′ end all contained an internal virion protein thought to be associated with puncturing the cell membrane in T7-like phages (*Mizuno, Ghai & Rodriguez-Valera, 2014*). Thus, it would appear that, for this shared genomic island at the population level, diversity occurs at the nucleotide level, rather than gene content level. The fact that a similar gene content has now been found in the Western English Channel (this study), the Sargasso Sea (*Zhao et al., 2013*) and the Mediterranean (*Mizuno, Ghai & Rodriguez-Valera, 2014*) may indicate this is a conserved feature across the HTVC010P-like phages. The encoding of a ribonuclease within a genomic island offers an interesting glimpse into the host-virus interactions that occur during infection and suggests that degradation of RNA is an important feature of the arms-race in HTVC010P-like phages with their *Pelagibacter* hosts. Whether this is to shut down host metabolism, or to hijack host metabolism through manipulation of regulatory machinery enriched in riboswitches (*Meyer et al., 2009*) requires further investigation.

In total, analysis of VirION reads using a strategy to combine short and long read assemblies (Fig. 2) generated 2,645 putative viral contigs >10 kbp from the Western English Channel. Of these, 2,279 were from the de-replicated short and hybrid De Bruijn Graph assemblies and 366 from polished long-read assemblies. Our dataset represents the first virome sequenced from the WEC and so we evaluated the global abundance of viral populations from the WEC by competitive mapping of 10 million subsampled short reads from both the WEC and the GOV dataset (*Roux et al., 2016a*). Representatives of viral populations from the WEC were then pooled with those >10 kb from the GOV dataset and other marine virome datasets (Table S2) to make a total dataset of 20,545 viral contigs. Following competitive read recruitment with FastViromeExplorer (*Tithi et al., 2018*), the top 50 most abundant viral genomes were identified in each of the WEC and GOV surface samples. Out of 1,598 contigs, 81 of the most abundant viral contigs were from long read assemblies of VirION reads from the WEC, representing a significant enrichment (hypergeometric test for enrichment, $p = 6.6 \times 10^{-19}$). WEC contigs from short-read only (42 contigs) and hybrid assemblies (77 contigs) were not significantly enriched in the most abundant viral contigs. Thus, it is likely that long-read assembly of VirION reads from the
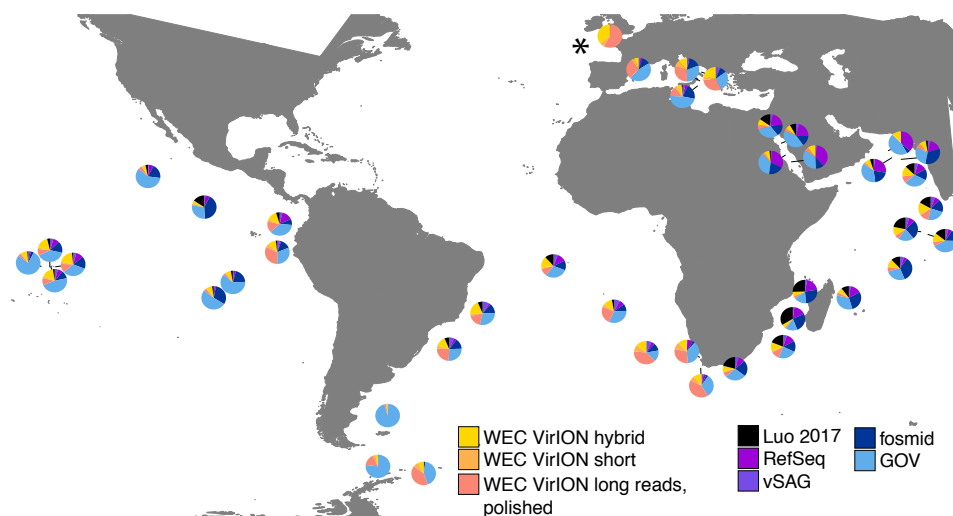
Figure 6 **VirION-derived viral genomes from the Western English Channel are abundant in global marine viromes.** Relative abundances were calculated via competitive recruitment of 10 million sub-sampled reads from each of 42 samples from the Global Ocean Virome (*Roux et al., 2016a*). Short reads were recruited against a database comprising VirION-derived viral genomes (both scaffolded and un-scaffolded De Bruijn Graph (DBG) assemblies and those from long-read assembly of VirION long reads) and viral genomes obtained from other key viral metagenomic studies (including those which have employed short-read sequencing ('GOV'; *Roux et al., 2016a*, and 'Luo 2017'; *Luo et al., 2017*), and long-sequence recovery via Single-virus genomics ('vSAG'; *Martinez-Hernandez et al., 2017*), and fosmid libraries ('fosmid'; *Mizuno et al., 2013*; *Mizuno et al., 2016*), and viruses from the NCBI RefSeq database v. 8.4 (all detailed in Table S2). The Western English Channel sample is indicated with a '*'.

Full-size ◘ DOI: 10.7717/peerj.6800/fig-6

WEC captured important and globally abundant viral taxa previously missed in the GOV datasets. Examination of relative abundance of WEC contigs in surface water samples from the GOV showed that contigs from long-read assemblies of VirION reads recruited a large proportion of the recruited reads from global samples, particularly in the Southern Atlantic Ocean and waters off the Western coasts of Southern Africa and South America (Fig. 6). In total, clustering VirION-derived contigs from the Western English Channel with contigs from previous studies (Table S2) by shared protein content produced 668 statistically supported viral clusters. Of these, 202 contained contigs derived from long-read assembly of VirION reads, but just three of these were comprised solely of these contigs. Thus, we are confident that previous findings suggesting viral diversity at the genera level in surface oceans has been largely documented (*Roux et al., 2016a*) are robust. Instead, we propose that long read assembly of VirION reads provides greater phylogenetic resolution of viral clusters by capturing members previously missed due to limitations in short-read assembly.

The most globally abundant and ubiquitous (identified in at least 10% of samples) viral genome was a contig from a hybrid assembly, denoted H_NODE_1248 (Fig. 7). This contig was 22.4 kbp in length and occupied a viral cluster (based on shared protein content) with 57 other members, including vSAG-37-F6 (9th most abundant ubiquitous virus and 13th most abundant across all samples), previously identified the most globally abundant virus (*Martinez-Hernandez et al., 2017*; *Martinez-Hernandez et al., 2018*). The viral cluster
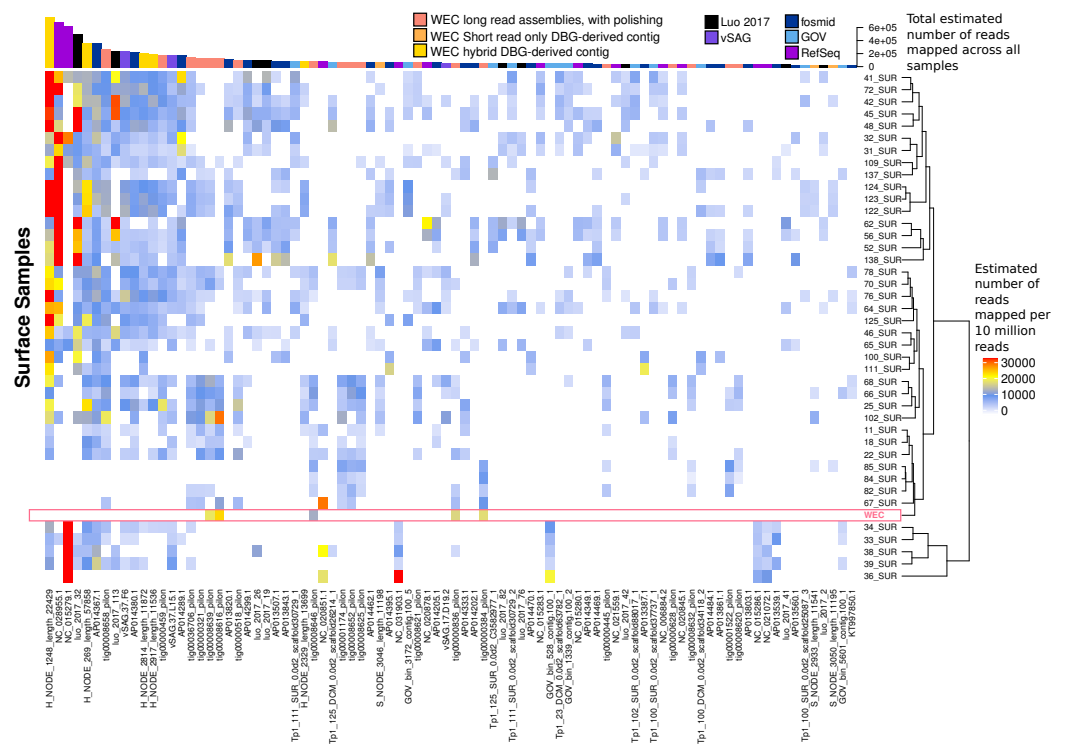
Figure 7 Ubiquity of VirION-derived Western English Channel viruses in Global Ocean surface waters. Heatmap shows the top 50 most abundant and ubiquitous (appear in >10% of samples) viral contigs in the surface samples of the Global Ocean Virome (*Roux et al., 2016a*). Competitive recruitment of 10 million subsampled short reads was performed using FastViromeExplorer (*Tithi et al., 2018*) against a contig database comprising: (1) viral population contigs from this study; (2) viral genomes derived from other key viral metagenomic studies (Table S2); (3) Viruses from the NCBI RefSeq database. Estimated abundances are calculated from the total number of reads mapped to a contig, with reads mapping to multiple contigs apportioned to a single contig via an expectation-maximum algorithm (*Tithi et al., 2018*). Matrix columns are ordered (left to right) by total number of mapped reads across all samples. The most abundant contig was H_NODE_1248, which is related at the genus level to the ubiquitous pelagiphage vSAG-37-F6. The Western English Channel sample is highlighted in a pink box, showing globally ubiquitous and abundant viruses from oceanic provinces were not particularly abundant in this coastal sample.

Full-size DOI: 10.7717/peerj.6800/fig-7

also contained 10 other contigs from long-read assembly of VirION reads, ranging in size from 10 kbp to 27 kbp. Interestingly, pelagiphage HTVC010P, once thought to be the most abundant virus on Earth (*Zhao et al., 2013*) was ranked 128th in global abundance and did not meet the criteria of being both ubiquitous (identified in at least 10% of the samples) and abundant (in the top 100 most abundant viral taxa for each sample). Upon its discovery as the most abundant global virus we previously urged a cautious interpretation as any representative of a new viral clade will recruit reads from all similar viruses in the environment (*Zhao et al., 2013*). As new representatives of these clades are captured in metagenomic data it is likely that competitive recruitment will split reads between all clade members, reducing the estimated abundance of any one single member.

60% of the top 50 most abundant populations in the WEC were represented by a WEC contig derived from long-read assemblies of VirION reads (Fig. S7). The viral community

in the WEC sample was dominated by a 39,972 bp circular genome from a hybrid assembly. Denoted H_NODE_525, this contig recruited 3.28 times more reads than the next most abundant contig (Fig. S7), but was not identified as globally abundant and ubiquitous (Fig. 7). This virus shared a viral cluster with the siphovirus *Pseudoalteromonas phage* vB_PspS-H6/1 but we were not able to determine its putative host despite using a variety of tools (*Ahlgren et al., 2016*; *Galiez et al., 2017*) (https://github.com/dutilh/CAT). A viral contig from hybrid assembly, denoted H_NODE_6 was the longest complete viral genome identified in this study, with a 316 kbp genome. In the short read-only assembly, this genome was broken into two contiguous contigs of 204 kbp and 112 kbp, respectively (Fig. S8). H_NODE_6 shared a viral cluster with the myoviruses *Cronobacter sakasakii* phage GAP32 and Enterobacter phage vB_KleM-RaK2. At 359 kb and 346 kb respectively (*Šimoliūnas et al., 2012*; *Abbasifar et al., 2014*), these are some of the largest phage genomes ever isolated. Recovery of this complete genome demonstrates the capacity for hybrid assembly with VirION reads to capture complete genomes of very large phages from complex communities on single contigs, which were fragmented using short-read only assemblies.

CONCLUSIONS

In summary, this investigation represents the first use of long-read sequencing for viral metagenomics. We have shown that using long-reads to scaffold short read De Bruijn Graph assemblies improves recovery of complete viral genomes. Furthermore, overlap-layout consensus assembly of VirION reads, followed by error correction with short reads captures abundant and ubiquitous viral populations that are missed (possibly as a result of genome fragmentation) by current short-read metagenomic methods. By combining these two approaches, our proposed bioinformatics pipeline maximises the capture of viral diversity whilst minimising the impact of high error rates associated with long-read sequencing and represents a major addition to the viral metagenomics toolset. Improved capture of viral genomic islands will enable better understanding of mechanisms underpinning host–virus interactions, as demonstrated in our capture of a shared genomic island on the newly observed HTVC010P-like pelagiphage tig404. 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, at a cost of <$1000 per flowcell. Thus, our approach represents a significant advantage in terms of cost, yield and efficiency over fosmid and single-amplified genome approaches to capturing marine viruses that are otherwise challenging to assemble.

As error rates associated with MinION technology continue to fall, we envisage less and less complementary short-read data being required for polishing. A recent update to basecalling methods has led to a significant reduction in indel errors and their associated impact on protein prediction (*Koren et al., 2019*). Furthermore, there is no technical reason to prevent our VirION approach being used in conjunction with PacBio sequencing to further reduce error rates using circular consensus sequencing. Such an approach would remove the need for short-read error correction (*Frank et al., 2016*) and avoid the remaining

indel errors observed following polishing of MinION read assemblies with short-read data. Reductions in DNA input requirements and/or improvements in DNA polymerases for increasing VirION amplicon lengths will further increase its utility in recovering viral genomes from metagenomic samples.

Ultimately, community efforts to align the input requirements of long-read sequencing with DNA recovery rates from viral communities will be rewarded by the ability to capture full-length viral genomes on single reads (*Houldcroft, Beale & Breuer, 2017*), including all associated nucleotide modifications (*Viehweger et al., 2018*). Oxford Nanopore sequencing interprets single stranded nucleotides as they pass through the pore and there are significant efforts to develop protocols for direct sequencing of RNA, dsDNA and ssDNA viruses (*Keller et al., 2018*; *Viehweger et al., 2018*; *McCabe et al., 2018*). It is theoretically possible that with the right combination of ligases and optimised buffers, we will soon be able to sequence dsDNA, ssDNA and RNA viruses, with associated nucleotide modifications, within a single library preparation. We expect that the VirION approach could be readily adapted for use with ssDNA or RNA viruses, provided appropriate amplification of starting material could be achieved to meet the requirements for sequencing. Therefore, the approach described here provides a significant step towards capturing the full diversity of the viral community.

VirION offers a framework for robust downstream bioinformatic approaches to maximise the benefits of long read sequencing, both now and as the technology continues to improve. Here, we have shown that VirION long-read metagenomics of dsDNA viral communities offers the potential to significantly improve our understanding of niche-differentiation, ecological speciation and the role of viruses in microbial communities within aquatic (*Roux et al., 2016a*) and soil (*Pratama & Van Elsas, 2018*) environments, human health (*Mirzaei & Maurice, 2017*; *Aggarwala, Liang & Bushman, 2017*) and industrial settings.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Competing Interests
The authors declare there are no competing interests.

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.
- Natalie Solonenko performed the experiments.
- Karen Moore performed the experiments, contributed reagents/materials/analysis tools.
- Lauren Chittick performed the experiments.
- Ann C. Gregory analyzed the data.
- 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.

DNA Deposition
The following information was supplied regarding the deposition of DNA sequences:
Sequencing data and assemblies are available at the European Nucleotide Archive under the project accession number PRJEB27181.

Data Availability
The following information was supplied regarding data availability:
All code and analyses can be found in a GitHub repository: https://github.com/btemperton/long_read_viromics.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6800#supplemental-information.

Abbasifar R, Griffiths MW, Sabour PM, Ackermann H-W, Vandersteegen K, Lavigne R, Noben J-P, Alanis Villa A, Abbasifar A, Nash JHE, Kropinski AM. 2014. Supersize me: *Cronobacter sakazakii* phage GAP32. *Virology* **460–461**:138–146 DOI 10.1016/j.virol.2014.05.003.

Aggarwala V, Liang G, Bushman FD. 2017. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA* **8**:12 DOI 10.1186/s13100-017-0095-y.

Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. 2016. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Research* **45**:39–53 DOI 10.1093/nar/gkw1002.

Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J. 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nature Biotechnology* **33**:296–300 DOI 10.1038/nbt.3103.

Bechhofer DH, Hue KK, Shub DA. 1994. An intron in the thymidylate synthase gene of Bacillus bacteriophage beta 22: evidence for independent evolution of a gene, its group I intron, and the intron open reading frame. *Proceedings of the National Academy of Sciences of the United States of America* **91(24)**:11669–11673 DOI 10.1073/pnas.91.24.11669.

Bolduc B, Jang HB, Doulcier G, You Z-Q, Roux S, Sullivan MB. 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**:e3243 DOI 10.7717/peerj.3243.

Bonilla N, Rojas MI, Netto G, Cruz F, Hung S, Rohwer F, Barr J. 2016. Phage on Tap—a quick and efficient protocol for the preparation of bacteriophage laboratory stocks. *PeerJ* **4**:e2261 DOI 10.7717/peerj.2261.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**:525–527 DOI 10.1038/nbt.3519.

Breitbart M, Thompson LR, Suttle CA, Sullivan MB. 2007. Exploring the vast diversity of marine viruses. *Oceanography* **20**:135–139 DOI 10.5670/oceanog.2007.58.

Brocchieri L, Karlin S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33(10)**:3390–3400 DOI 10.1093/nar/gki615.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, De Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans C, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348(6237)**:1261498 DOI 10.1126/science.1261498.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Chu FK, Maley GF, Maley F, Belfort M. 1984. Intervening sequence in the thymidylate synthase gene of bacteriophage T4. *Proceedings of the National Academy of Sciences of the United States of America* **81**:3049–3053 DOI 10.1073/pnas.81.10.3049.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. 2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311(5768)**:1768–1770 DOI 10.1126/science.1122050.

Cumming G. 2014. The new statistics: why and how. *Psychological Science* **25**:7–29 DOI 10.1177/0956797613504966.

Cunningham BR, Brum JR, Schwenck SM, Sullivan MB, John SG. 2015. An inexpensive, accurate, and precise wet-mount method for enumerating aquatic viruses. *Applied and Environmental Microbiology* **81**:2995–3000 DOI 10.1128/AEM.03642-14.

Dabney A, Storey JD, Warnes GR. 2015. qvalue: Q-value estimation for false discovery rate control. R package version 2.12.0. *Available at https://www.bioconductor.org/packages/release/bioc/html/qvalue.html*.

Delcher AL, Salzberg SL, Phillippy AM. 2003. Using MUMmer to identify similar regions in large sequence sets. *Current Protocols in Bioinformatics* **Chapter 10**: Unit 10.3.

Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB. 2014. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513**:242–245 DOI 10.1038/nature13459.

Driscoll CB, Otten TG, Brown NM, Dreher TW. 2017. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Standards in Genomic Sciences* **12**:9 DOI 10.1186/s40793-017-0224-8.

Duhaime MB, Deng L, Poulos BT, Sullivan MB. 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environmental Microbiology* **14**:2526–2537 DOI 10.1111/j.1462-2920.2012.02791.x.

Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M. 2013. A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evolutionary Biology* **13**:33 DOI 10.1186/1471-2148-13-33.

Forterre P. 2013. The virocell concept and environmental microbiology. *The ISME Journal* **7**:233–236 DOI 10.1038/ismej.2012.110.

Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, Pope PB. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports* **6**:25373 DOI 10.1038/srep25373.

Galiez C, Siebert M, Enault F, Vincent J, Söding J. 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**:3113–3114 DOI 10.1093/bioinformatics/btx383.

Giovannoni S. 2017. SAR11 bacteria: the most abundant plankton in the oceans. *Annual Review of Marine Science* **9**:231–255 DOI 10.1146/annurev-marine-010814-015934.

Houldcroft CJ, Beale MA, Breuer J. 2017. Clinical and biological insights from viral genome sequencing. *Nature Reviews. Microbiology* **15**:183–192 DOI 10.1038/nrmicro.2016.182.

Hurwitz BL, Deng L, Poulos BT, Sullivan MB. 2013. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology* **15**:1428–1440 DOI 10.1111/j.1462-2920.2012.02836.x.

Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biology* **14**:R123 DOI 10.1186/gb-2013-14-11-r123.

Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLOS ONE* **8(2)**:e57355 DOI 10.1371/journal.pone.0057355.

Hurwitz BL, U'Ren JM. 2016. Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology* **31**:161–168 DOI 10.1016/j.mib.2016.04.002.

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M. 2015. Improved data analysis for the MinION nanopore sequencer. *Nature Methods* **12**:351–356 DOI 10.1038/nmeth.3290.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**:338–345 DOI 10.1038/nbt.4060.

John SG, Mendez CB, Deng L, Poulos B, Kauffman AK, Kern S, Brum J, Polz MF, Boyle EA, Sullivan MB. 2011. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environmental Microbiology Reports* **3**:195–202 DOI 10.1111/j.1758-2229.2010.00208.x.

Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ, Neuhaus EB, Dugan VG, Wentworth DE, Barnes JR. 2018. Direct RNA sequencing of the coding complete influenza a virus genome. *Scientific Reports* **8**:14408 DOI 10.1038/s41598-018-32615-8.

Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology* **23**:110–120 DOI 10.1016/j.mib.2014.11.014.

Koren S, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2019. Reply to errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology* **37(2)**:127–128 DOI 10.1038/s41587-018-0005-y.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**:722–736 DOI 10.1101/gr.215087.116.

Krishnamurthy SR, Wang D. 2017. Origins and challenges of viral dark matter. *Virus Research* **239**:136–142 DOI 10.1016/j.virusres.2017.02.002.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359 DOI 10.1038/nmeth.1923.

Laver TW, Caswell RC, Moore KA, Poschmann J, Johnson MB, Owens MM, Ellard S, Paszkiewicz KH, Weedon MN. 2016. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. *Scientific Reports* **6**:21746 DOI 10.1038/srep21746.

Loman NJ, Quick J, Simpson JT. 2015. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12**:733–735 DOI 10.1038/nmeth.3444.

Luo E, Aylward FO, Mende DR, DeLong EF. 2017. Bacteriophage distributions and temporal variability in the ocean's interior. *mBio* **8(6)**:e01903-17 DOI 10.1128/mBio.01903-17.

Mahmoudabadi G, Phillips R. 2018. A comprehensive and quantitative exploration of thousands of viral genomes. *eLife* **7**:e31955 DOI 10.7554/eLife.31955.

Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, De la Cruz Peña MJ, Martínez JM, Anton J, Gasol JM, Rosselli R, Rodriguez-Valera F, Sullivan MB, Acinas SG, Martinez-Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature Communications* **8**:15892 DOI 10.1038/ncomms15892.

Martinez-Hernandez F, Fornas Ò, Lluesma Gomez M, Garcia-Heredia I, Maestre-Carballa L, López-Pérez M, Haro-Moreno JM, Rodriguez-Valera F, Martinez-Garcia M. 2018. Single-cell genomics uncover *Pelagibacter* as the putative host of the extremely abundant uncultured 37-F6 viral population in the ocean. *The ISME Journal* **13**:232–236 DOI 10.1038/s41396-018-0278-7.

McCabe M, Cormican P, Johnston D, Earley B. 2018. Simultaneous detection of DNA and RNA virus species involved in bovine respiratory disease by PCR-free rapid tagmentation-based library preparation and MinION nanopore sequencing. *bioRxiv* DOI 10.1101/269936.

Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, Antin P. 2016. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLOS Biology* **14(1)**:e1002342 DOI 10.1371/journal.pbio.1002342.

Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, Breaker RR. 2009. Identification of candidate structured RNAs in the marine organism *Candidatus* Pelagibacter ubique. *BMC Genomics* **10**:268 DOI 10.1186/1471-2164-10-268.

Mirzaei MK, Maurice CF. 2017. Ménage à trois in the human gut: interactions between host, bacteria and phages. *Nature Reviews. Microbiology* **15(7)**:397–408 DOI 10.1038/nrmicro.2017.30.

Mizuno CM, Ghai R, Rodriguez-Valera F. 2014. Evidence for metaviromic islands in marine phages. *Frontiers in Microbiology* **5(27)** DOI 10.3389/fmicb.2014.00027.

Mizuno CM, Ghai R, Saghaï A, López-García P, Rodriguez-Valera F. 2016. Genomes of abundant and widespread viruses from the deep ocean. *mBio* **7(4)**:e00805-16 DOI 10.1128/mBio.00805-16.

Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. *PLOS Genetics* **9(12)**:e1003987 DOI 10.1371/journal.pgen.1003987.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**:5269–5273 DOI 10.1073/pnas.76.10.5269.

Nepusz T, Yu H, Paccanaro A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods* **9**:471–472 DOI 10.1038/nmeth.1938.

Noble RT. 2001. Enumeration of viruses. In: *Methods in microbiology.* Cambridge: Academic Press, 43–51.

Noguchi H, Taniguchi T, Itoh T. 2008. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Research* **15(6)**:387–396 DOI 10.1093/dnares/dsn027.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**:824–834 DOI 10.1101/gr.213959.116.

Ohio Supercomputer Center. 1987. *Ohio supercomputer center.* Columbus: Ohio Supercomputer Center.

Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, Pop M. 2017. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics* Epub ahead of print Aug 7 2017 DOI 10.1093/bib/bbx098.

Pratama AA, Van Elsas JD. 2018. The neglected soil virome—potential role and impact. *Trends in Microbiology* **26**:649–662 DOI 10.1016/j.tim.2017.12.004.

Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solo-nenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaque D, Oceans CTara, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016a. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**:689–693 DOI 10.1038/nature19366.

Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**:e3817 DOI 10.7717/peerj.3817.

Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**:e985 DOI 10.7717/peerj.985.

Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, Cole-man ML, Breitbart M, Sullivan MB. 2016b. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**:e2777 DOI 10.7717/peerj.2777.

Shagin DA, Lukyanov KA, Vagner LL, Matz MV. 1999. Regulation of average length of complex PCR product. *Nucleic Acids Research* **27**:e23 DOI 10.1093/nar/27.18.e23.

Šimoliūnas E, Kaliniene L, Truncaite L, Klausa V, Zajančkauskaite A, Meškys R. 2012. Genome of *Klebsiella sp.*-infecting bacteriophage vB_KleM_RaK2. *Journal of Virology* **86**:5406 DOI 10.1128/JVI.00347-12.

Sullivan MB. 2015. Viromes, not gene markers, for studying double-stranded DNA virus communities. *Journal of Virology* **89**:2459–2461 DOI 10.1128/JVI.03289-14.

Suttle CA. 2005. Viruses in the sea. *Nature* **437**:356–361 DOI 10.1038/nature04160.

Suttle CA. 2007. Marine viruses—major players in the global ecosystem. *Nature Reviews. Microbiology* **5**:801–812 DOI 10.1038/nrmicro1750.

Temperton B, Giovannoni SJ. 2012. Metagenomics: microbial diversity through a scratched lens. *Current Opinion in Microbiology* **15(5)**:605–612 DOI 10.1016/j.mib.2012.07.001.

Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vázquez-Baeza Y, González A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J, Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G, Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA, Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA, Knight R, Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**:457–463 DOI 10.1038/nature24621.

Tithi SS, Aylward FO, Jensen RV, Zhang L. 2018. FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data. *PeerJ* **6**:e4227 DOI 10.7717/peerj.4227.

Torsvik V, Øvreås L. 2002. Microbial diversity and function in soil: from genes to ecosystems. *Current Opinion in Microbiology* **5(1)**:240–245 DOI 10.1016/S1369-5274(02)00324-7.

Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA, Giovannoni SJ. 2009. Seasonality and vertical structure of microbial communities in an ocean gyre. *The ISME Journal* **3**:1148–1163 DOI 10.1038/ismej.2009.60.

Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz M. 2018. Nanopore direct RNA sequencing reveals modification in full-length coronavirus genomes. *bioRxiv preprint* DOI 10.1101/483693.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE* **9(11)**:e112963 DOI 10.1371/journal.pone.0112963.

Warr A, Watson M. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology* **37**:124–126 DOI 10.1038/s41587-018-0004-z.

Weinbauer MG. 2004. Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* **28(2)**:127–181 DOI 10.1016/j.femsre.2003.08.001.

Weirather JL, De Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, Buck D, Au KF. 2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore

Technologies and their applications to transcriptome analysis. *F1000Research* **6**:100 DOI 10.12688/f1000research.10571.2.

Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* **13(6)**:e1005595 DOI 10.1371/journal.pcbi.1005595.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, Ellisman M, Deerinck T, Sullivan MB, Giovannoni SJ. 2013. Abundant SAR11 viruses in the ocean. *Nature* **494**:357–360 DOI 10.1038/nature11921.