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], **Ben Temperton** [Corresp. 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, Ohio, United States

4 Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, Ohio, United States

Corresponding Author: Ben Temperton
Email address: b.temperton@exeter.ac.uk

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.

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

4   **Authors**
5   Joanna Warwick-Dugdale[1,2], Natalie Solonenko[3], Karen Moore[2], Lauren Chittick[3], Ann C.
6   Gregory[3], Michael J. Allen[1,2], Matthew B. Sullivan[3,4], Ben Temperton[2].
7
8   [1.] Plymouth Marine Laboratory, Plymouth, UK
9   [2.] School of Biosciences, University of Exeter, Exeter, UK.
10 [3.] Department of Microbiology, and [4] Civil, Environmental and Geodetic Engineering, The Ohio
11 State University, Ohio, USA.
12
13 Corresponding Author:
14 Ben Temperton
15 email: b.temperton@exeter.ac.uk

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

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

25 [1.] Plymouth Marine Laboratory, Plymouth, UK
26 [2.] School of Biosciences, University of Exeter, Exeter, UK.
27 [3.] Department of Microbiology, and [4] Civil, Environmental and Geodetic Engineering, The Ohio
28 State University, Ohio, USA.

30 Corresponding Author:
31 Ben Temperton
32 email: b.temperton@exeter.ac.uk

34 **Abstract**
35 Marine viruses impact global biogeochemical cycles via their influence on host community
36 structure and function, yet our understanding of viral ecology is constrained by limitations in host
37 culturing and a lack of reference genomes and 'universal' gene markers to facilitate community
38 surveys. Short-read viral metagenomic studies have provided clues to viral function and first
39 estimates of global viral gene abundance and distribution, but their assemblies are confounded
40 by populations with high levels of strain evenness and nucleotide diversity (microdiversity),
41 limiting assembly of some of the most abundant viruses on Earth. Such features also challenge
42 assembly across genomic islands containing niche-defining genes that drive ecological
43 speciation. These populations and features may be successfully captured by single-virus
44 genomics and fosmid-based approaches, at least in abundant taxa, but at considerable cost and
45 technical expertise. Here we established a low-cost, low-input, high throughput alternative
46 sequencing and informatics workflow to improve viral metagenomic assemblies using short-read
47 and long-read technology. The 'VirION' (Viral, long-read metagenomics via MinION sequencing)
48 approach was first validated using mock communities where it was found to be as relatively
49 quantitative as short-read methods and provided significant improvements in recovery of viral
50 genomes. We then then applied VirION to the first metagenome from a natural viral community
51 from the Western English Channel. In comparison to a short-read only approach, VirION: (i)
52 increased number and completeness of assembled viral genomes; (ii) captured abundant,
53 highly microdiverse virus populations, and (iii) captured more and longer genomic islands.
54 Together, these findings suggest that VirION provides a high throughput and cost-effective
55 alternative to fosmid and single-virus genomic approaches to more comprehensively explore
56 viral communities in nature.
57

58 **Introduction**
59 The marine bacterial communities that regulate global carbon biogeochemical cycles are
60 themselves structured by selective, phage-mediated lysis (Weinbauer, 2004; Suttle, 2007).

61   Bacteria co-evolve with their phages and exchange genetic information, and phages even
62   'reprogram' hosts during infection so as to channel host metabolism towards phage replication
63   (Forterre, 2013; Hurwitz, Hallam & Sullivan, 2013; Hurwitz & U'Ren, 2016). Over the last
64   decade, the convergence of high throughput sequencing and the use of universal taxonomic
65   marker genes for bacteria have revolutionised our understanding of microbial ecology (Torsvik &
66   Øvreås, 2002; Treusch et al., 2009; Thompson et al., 2017). Problematically, however, viral
67   ecologists lack parallel approaches. First, PCR-amplified marker genes are limited to a narrow
68   subset of the viral community, and require degeneracies and amplification conditions that
69   undermine the quantitative nature of the data (Sullivan, 2015). Second, while short-read viral
70   metagenomics studies to date have provided clues to viral function (e.g. virally encoded, host-
71   derived central metabolism genes (known as Auxiliary Metabolic Genes: AMGs) (Breitbart et al.,
72   2007; Hurwitz, Hallam & Sullivan, 2013), and first estimates of global viral gene abundance and
73   distribution (Brum et al., 2015; Roux et al., 2016a), they suffer from technical limitations. This is
74   because short-read assemblies are composites of populations 'features' (Mizuno, Ghai &
75   Rodriguez-Valera, 2014), with successful assembly a function of coverage and branch
76   resolution in assembly graphs (Temperton & Giovannoni, 2012; Olson et al., 2017). This limits
77   our ability to assemble viral populations where multiple strains are abundant and microdiverse
78   (Roux et al., 2017), as well as genomic regions of high diversity, such as genomic islands (GIs),
79   which, in microbes, often contain niche-defining genes that drive ecological speciation (Coleman
80   et al., 2006). In these latter regions, assembly is impeded by low coverage and/or repeat
81   regions at the boundaries (Mizuno, Ghai & Rodriguez-Valera, 2014; Ashton et al., 2015).
82
83   These are not just technical limitations - emerging data suggests that these obstacles alter our
84   understanding of viral roles on important taxa and global carbon biogeochemistry. For example,
85   the globally dominant members of the chemoheterotrophic order Pelagibacterales comprise up
86   to 25% of all bacterioplankton and are major contributors in the conversion of marine dissolved
87   organic matter back to atmospheric CO_2 (Giovannoni, 2017). Their associated viruses dominate
88   global oceans (Zhao et al., 2013; Martinez-Hernandez et al., 2018) and are likely to contribute
89   significantly to carbon turnover in surface water by release of labile intracellular carbon during
90   lysis (Suttle, 2005, 2007). However, the genomes of viruses associated with Pelagibacterales
91   contain numerous GIs and/or high microdiversity (Zhao et al., 2013; Martinez-Hernandez et al.,
92   2018). Such features fragment genomes in short-read assemblies, which reduces
93   representation following contig size-selection for downstream analyses (Martinez-Hernandez et
94   al., 2017; Roux et al., 2017). Though single-virus genomics (Martinez-Hernandez et al., 2017)
95   and fosmid-based approaches (Mizuno et al., 2013, 2016) can overcome such issues, these
96   methods are technically challenging and costly to implement.
97
98   Alternatively, recent advances in long-read sequencing technology might be leveraged to better
99   capture microdiverse viral populations and genomic islands. Such approaches can yield very
100 long reads (>800 kbp) (Jain et al., 2015, 2018; Loman, Quick & Simpson, 2015), which would
101 be long enough to capture complete genomes of double-stranded DNA bacteriophages
102 ('phages') (10-617.5 kbp (Mahmoudabadi & Phillips, 2018)). At a minimum, such long reads
103 could span genomic global- and local repeat regions, which tangle the De Bruijn Graph and
104 fragment the assembly (Koren & Phillippy, 2015). Long reads may also overcome assembly

105 challenges in regions of low coverage, to improve overall assembly of genomes from both
106 cultured isolates (Wick et al., 2017) and metagenomics (Frank et al., 2016; Driscoll et al., 2017).
107 It is also probable that long-read assemblies using overlap-layout-consensus would be less
108 prone to microdiversity-associated fragmentation of genomes observed in De Bruijn Graph
109 approaches (Martinez-Hernandez et al., 2017; Roux et al., 2017).
110
111 The challenge is that long-read technologies (both from PacBio and Oxford Nanopore) currently
112 require large amounts of input DNA (micrograms (Jain et al., 2018) instead of nanograms
113 commonly available from natural viral communities in seawater (Hurwitz et al., 2013)).
114 Furthermore, PacBio subreads and nanopore reads have high error rates (5-10%), with the
115 former enriched in insertion errors and the latter enriched in insertion-deletion errors (Weirather
116 et al., 2017). Indel errors shift the reading frame of the DNA sequence and confound gene-
117 calling algorithms, artificially inflating the number of identified stop codons and producing shorter
118 gene calls (Warr & Watson, 2019). This is a particular problem for viral metagenomics as the
119 median length of genes in dsDNA phages is approximately half that of their bacterial hosts (408
120 bp vs 801 bp, respectively) (Brocchieri & Karlin, 2005; Mahmoudabadi & Phillips, 2018), and the
121 vast majority of viral genes in both dsDNA viral isolates and viral metagenomes (>50% and up
122 to 93%, respectively) have no known function (Hurwitz & Sullivan, 2013; Mahmoudabadi &
123 Phillips, 2018), making it difficult to evaluate the quality of gene calls from metagenomic
124 assemblies.
125
126 Here, we adapted a Long-Read Linker-Amplified Shotgun Library (LASL) approach for
127 quantitative viral metagenomics (Duhaime et al., 2012) to obtain sufficient quantities of high-
128 molecular weight DNA from nanograms of viral community dsDNA for sequencing using the
129 MinION sequencer from Oxford Nanopore Technology. We then established a bioinformatic
130 workflow to combine such long-read data with complimentary short-read sequencing data to
131 maximise the advantages and minimise the weaknesses of both sequencing technologies.
132 Following validation on mock viral communities, we applied our new approach to the first marine
133 viral metagenome from the Western English Channel (WEC). Here, we present the first use of
134 long-read sequencing technology for viral metagenomics and show that this novel approach
135 provides significant benefits when combined with short-read metagenomics.
136
137 **Materials & Methods**
138
139 ***Construction of the mock viral community:*** A mock viral community comprised of six isolated
140 and sequenced marine Caudovirales with genome sizes ranging from 38.5-129.4 kbp was
141 produced as described previously (Roux et al., 2016b). Briefly, viruses were cultivated from host
142 *Pseudoalteromonas* or *Cellulophaga* via plaque assay, collected into MSM buffer (0.45 M NaCl,
143 0.05 M Mg, 0.05 M Tris base, pH 7.6) and purified by 0.2 µm filtration followed by treatment with
144 DNase I (100 U/mL for 2 hr at RT; terminated by the addition of 0.1 M EGTA and 0.1 M EDTA).
145 Viral capsids were enumerated via epifluorescence microscopy (SYBR Gold; wet mount
146 method) (Noble, 2001; Cunningham et al., 2015). 1.4×10^9 virus particles from each culture
147 were pooled, and DNA extracted via the Wizard® DNA Clean-up System (Promega A7280).
148 DNA was quantified via Qubit fluorometer (Life Technologies).

149
150   ***Construction of the Western English Channel viral metagenome:*** 20 L of seawater was
151   collected in rosette-mounted Niskin bottles at a depth of 5m from the Western Channel
152   Observatory (WCO; http://www.westernchannelobservatory.org.uk/) coastal station 'L4'
153   (50°15.0'N; 4°13.0'W) on the 28th September 2016. Seawater was transferred immediately to a
154   clean collection bottle, and processed to remove the cellular fraction (within 4 hours of
155   collection) via sequential filtration through glass fibre (GF/D: pore size 2.7 μm) and
156   polyethersulfone (pore size 0.22 μm) filters in a 142 mm polycarbonate rig, with peristaltic pump.
157   Precipitation of viruses from filtrate (denoted as the viral fraction) and primary concentration of
158   virus particles was conducted by iron chloride flocculation and collection on 1.0 μm
159   polycarbonate filters (John et al., 2011) and stored in the dark at 4°C. Viruses were
160   resuspended in ascorbate-EDTA buffer (0.1 M EDTA, 0.2 M $MgCl_2$, 0.2 M ascorbic acid, pH
161   6.1), and transferred to Amicon Ultra 100 kDa centrifugal filter units (Millipore UFC910024)
162   (Hurwitz et al., 2013) that had been pre-treated with 1 % bovine serum albumin buffer to
163   minimise capsid-filter adhesion (Deng et al., 2014) and flushed with SM buffer (0.1 M NaCl; 0.05
164   M Tris-HCl; 0.008 M $MgCl_2$). Following concentration to 500-600 μL, virus particles were
165   washed with SM buffer (39) and purified with DNase I (100 U/mL; 2 hr at RT) to remove
166   unprotected DNA (i.e. encapsulated DNA); DNase I activity was terminated by the addition of
167   0.1 M EGTA and 0.1 M EDTA (Hurwitz et al., 2013). Viral DNA was extracted from concentrated
168   and purified viral particles using the Wizard® DNA Clean-up System (Promega A7280),
169   removing PCR inhibitors (John et al., 2011).
170
171   ***Library preparation, amplification and sequencing:*** For short-read sequencing, Illumina
172   libraries were generated from 1 ng of either mock viral community DNA (Table S1), or 1 ng of
173   environmental viral-fraction DNA, using Nextera XT v2 kits (Illumina) and the manufacturer's
174   protocol. After 12 cycles of amplification, the concentration and distribution in fragment sizes of
175   the Illumina libraries were determined via Qubit and Bioanalyzer (Agilent), respectively. DNA
176   was sequenced as 2 × 300 bp paired-end sequence reads, on a HiSeq 2500 (Illumina Inc.) in
177   rapid mode, by the Exeter Sequencing Service (University of Exeter, UK).
178
179   For VirION libraries (Figure 1), 20 ng (mock viral community) or 100 ng (WEC viral-fraction) of
180   DNA was sheared to fragments averaging 8 kbp length via g-TUBE (Covaris 520079) as
181   required to optimise MinION flow cell sequencing efficiency/yield (Oxford Nanopore
182   Technologies: ONT). End-repair of DNA fragments, amplification of DNA with PCR-adapter
183   ligation (i.e. Linker Amplified Shotgun Library: LASL preparation), and preparation of MinION-
184   compatible libraries was performed following the manufacturer's protocols for "2D Low input
185   genomic DNA with PCR" using the 'Ligation Sequencing kit 2D' (ONT SQK-LSK208). PCR
186   reaction conditions were modified with reference to NEBNext manufacturer's instructions in
187   order to maximise DNA yield, whilst minimising production of chimeric sequences, as follows: 3
188   mins at 95°C (initial denaturation), 15 cycles of: 15 secs at 95°C (denaturation), 15 secs at 62°C
189   (annealing), 5 min at 72°C (extension); finally, 5 min at 72°C (final extension)) followed by 0.4 ×
190   AMPure bead clean-up). ~1.5 μg of end-repaired, amplified DNA was carried forward for
191   sequencing adapter ligation followed by purification of adapted DNA using MyOne C1
192   Streptavidin beads (Thermo Fisher Scientific Inc. 65001). The prepared long read library was

193 sequenced on a single MinION Mk 1B flow cell with R9.4 pore chemistry for 48 hours (Note - to
194 remain up to date with changing ONT chemistry, a 1D ligation version of this protocol has also
195 been tested and is available on protocols.io (https://www.protocols.io/view/virion-long-read-low-
196 input-viral-metagenomic-sequ-p8fdrtn)). Quality control of short and long read libraries was
197 performed as described in Supplementary Methods. Hiqh quality sequence data were used to
198 generate short-read De Bruijn Graph assemblies (using metaSPAdes v. 3.11 (Nurk *et al.*,
199 2017)), hybrid long-read scaffolded De Bruijn Graph assemblies (using metaSPAdes, with --
200 nanopore parameter), and long-read overlap-layout consensus assemblies (with Canu (Koren et
201 al., 2017) following optimisation for metagenomic data) (see Supplementary Methods). Rates of
202 chimerism in both VirION reads (formed during PCR amplification) and assemblies (formed by
203 mis-assembly) were evaluated by aligning reads and contigs, respectively, from the mock viral
204 community to their associated genomes (Table S1; Supplementary Methods).

205
206
207
208
209

210 ***Maximizing the benefits of long read and short read assemblies:*** We developed a
211 bioinformatic pipeline to maximise the benefits of VirION reads for viral metagenomics (Figure
212 2). Briefly, long-read assembly contigs of VirION reads were 'polished' with matching short-read
213 sequences to remove sequencing error via consensus base-calling (using Pilon (Walker et al.,
214 2014) v1.22). In order to capture the longest assemblies available from the short-read data,
215 scaffolds from short-read and hybrid assemblies were combined and dereplicated using a cut-
216 off of 95% average nucleotide identity over 80% of the length (via MUMmer v3.23 (Delcher,
217 Salzberg & Phillippy, 2003) to cluster highly similar contigs into viral populations (Roux et al.,
218 2016a). The longest representatives of each population were carried forward for analysis.
219 Population representatives > 10 kbp were pooled with polished long-read assembly contigs >
220 10kbp and evaluated with VirSorter (Roux et al., 2015) (in virome decontamination mode) to
221 identify putative viral contigs. Reads classified as category 3 (deemed unusual, but not
222 necessarily viral (Roux et al., 2015)) were excluded from downstream analyses. Circular contigs
223 (i.e. where the contig has matching ends) were identified by VirSorter and used as a proxy for
224 successful assembly of a complete genome. Matching short-read data was then mapped
225 against the representative viral population sequences (using bowtie2 (Langmead & Salzberg,
226 2012)) for use in evaluating (1) relative abundances of contigs; (2) whether long read assembly
227 captured more microdiverse genomes; and (3) recovery of genomic islands and their predicted
228 functional composition at the population level (*see Supplementary Methods*).

229

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

237 decontamination mode and the number of scaffolds classified as viral were calculated for each
238 replicated assembly. Statistical significance of the number of viral or circular viral contigs
239 between hybrid and short-read assemblies was calculated by a two-sided Student t-test
240 between triplicate replicates at each sequencing depth.
241
242 ***Validation of error correction of long reads in viral metagenomic data.*** We evaluated
243 whether it was possible to use short-read data to correct base-calling errors in long-read
244 environmental metagenomic data in a similar way to that used for genomes of bacteria and
245 eukaryotes from axenic samples (Walker et al., 2014). Triplicate subsamples of short-read
246 sequence data from Western English Channel at different sequencing depths were mapped
247 against the long-read assemblies of VirION reads. For error-correction with Pilon (Walker et al.,
248 2014) and median coverage, total number of fixed deletions and fixed insertions at each
249 coverage depth were calculated. We then evaluated whether error-correction could be used to
250 reduce the impact of frameshift errors on predicted gene length. Predicted coding sequences
251 were identified using MetaGeneAnnotator (Noguchi, Taniguchi & Itoh, 2008) on the following:
252 (1) uncorrected VirION reads; (2) long-read assemblies of VirION reads; (3) long-read
253 assemblies of VirION reads polished with the full short-read dataset; (4) contigs from scaffolded
254 short-read assemblies; (5) contigs from the hybrid assembly. Distributions of the lengths of
255 predicted coding sequences were compared against those in the genomes of *Caudovirales* from
256 the NCBI RefSeq database (v.8.4), predicted proteins from the GOV (Roux et al., 2016a) and
257 the single-amplified viruses in Martinez-Hernandez *et al*. (2017). Effect size of different
258 assembly types on genomic island length and density and associated 95% confidence intervals
259 (CI) were calculated from bootstrapped medians (Cumming, 2014). For each bootstrap, 1000
260 predicted proteins were randomly subsampled from each dataset and their median length was
261 calculated.
262
263 ***Analysis of Tig404 - a contig closely related to Pelagiphage HTVC010P:*** Phage contigs
264 closely related to Pelagiphage HTVC010P were identified using vContact2 to cluster contigs at
265 the ICTV-accepted level of genera by shared gene content (Bolduc et al., 2017). Within this viral
266 cluster, contig *tig404*, from polished long-read assembly of VirION reads, was identified as a
267 circular viral contig by VirSorter (Roux et al., 2015). Whole genome alignment was performed
268 with MUmmer (Delcher, Salzberg & Phillippy, 2003) to calculate average nucleotide identity to
269 HTVC010P. Contigs from short-read only and hybrid assemblies that shared 95% nucleotide
270 identity over 80% of their length to tig404 were identified and mapped back to their respective
271 loci with MUmmer (Delcher, Salzberg & Phillippy, 2003). Genomic islands and nucleotide
272 diversity of tig404 were calculated as described previously. To evaluate the contents of a 5.3kb
273 genomic island, unpolished VirION reads were mapped back against the tig404 genome and
274 those which mapped to at least 100 bp on the borders of the genomic island were extracted.
275 Mapped reads extending at least 1 kb into the genomic island were used as a query in a
276 tBLASTx best-BLAST (Camacho et al., 2009) search against the NCBI NR database to
277 annotate the reads whilst minimising the adverse impact of sequencing error within the
278 uncorrected reads.
279
280 ***Estimating relative abundance and viral clusters of WEC viruses in viral metagenomes:***

281 FastViromeExplorer (Tithi et al., 2018) v.1.1 was used to quantify the relative abundances of
282 WEC viral contigs. FastViromeExplorer is built upon the Kallisto (Bray et al., 2016) framework
283 and competitively recruits reads against contigs, allowing for accurate recruitment to contigs that
284 may share a degree of sequence similarity. Briefly, high quality short read datasets from the
285 Global Ocean Virome (Roux et al., 2016a) and from our Western English Channel sample were
286 randomly subsampled to 10 million reads using seqtk (https://github.com/lh3/seqtk) to
287 standardise per-sample sequencing effort, with the number of reads selected to balance
288 detection of lower abundance viral populations with maximising the number of samples that
289 could be included in the survey. Subsampled reads for each sample were recruited against a
290 Kallisto index comprising (1) The viral genomes >10 kbp identified in this study; (2) A selection
291 of phage genomes >10 kbp from key metagenomic studies (Roux et al., 2016a; Martinez-
292 Hernandez et al., 2017; Luo et al., 2017); (3) Cultured viruses from the NCBI RefSeq viral
293 database (v8.4) (Table S2). Contigs >10 kbp were selected to maximise accuracy of VirSorter to
294 correctly identify viral contigs (Roux et al., 2015). For inclusion in downstream abundance
295 analyses, contigs with less than 40% coverage as calculated by FastViromeExplorer were
296 classified as having zero abundance to avoid over-representation of partial matches (Tithi et al.,
297 2018). The top 100 most abundant contigs from each sample were also selected for
298 downstream analyses. All phage genomes > 10kbp (including those from RefSeq) were
299 processed using VirSorter (v.1.03) on the CyVerse cyberinfrastructure (Merchant et al., 2016) to
300 standardise gene-calling prior to clustering of viruses into ICTV-recognised genera by shared
301 gene content using vContact2 (Bolduc et al., 2017). In the final stage of clustering, vContact2
302 uses ClusterONE (Nepusz, Yu & Paccanaro, 2012) and assigns a p-value to a cluster
303 depending on whether the in-cluster edge weights are significantly higher than the out-cluster
304 edge weights. Q-values were calculated from cluster p-values using the qvalue R package
305 (Dabney, Storey & Warnes, 2015) to account for multiple testing and a q-value cutoff of <0.05
306 was used to identify statistically significant clusters.
307
308 **Results & Discussion**
309
310 Here, we present the first use of long-read sequencing technology for viral metagenomics and
311 show that this novel approach provides significant benefits when combined with short-read
312 metagenomics. Our bioinformatics pipeline overcame the high sequencing error associated with
313 long-read technology and the addition of long reads enabled capture of complete viral genomes
314 which were globally ubiquitous, and not represented by short-read only assemblies. Long-read
315 assemblies also significantly improved the capture of viral genomic islands, demonstrating that
316 this advance will facilitate better understanding of niche-differentiation and ecological speciation
317 of viruses in environmental samples.
318
319 *Assembly of VirION reads successfully captured mock viral community genomes and retained*
320 *relative abundance information:* VirION sequencing of the mock viral community produced
321 359,338 high quality (Q>10) long reads (median length: 4,099 bp; max length 18,644 bp). 95%
322 of the reads (341,718) mapped back to the genomes of the mock viral community. Considering
323 viral DNA was sheared to 6-8 kbp fragments, the length of amplicons following LASL were
324 shorter than expected, presumably due to preferential PCR amplification of shorter fragments

325   (Shagin et al., 1999) (Figure S1a) and preferential diffusion (and thus sequencing) of shorter
326   reads within the flowcell microfluidics (Figure S1b). Only 0.95% of LASL amplified reads were
327   classified as chimeric (mapping to more than one location of the same or different genomes of
328   the mock viral community), suggesting 15 rounds of PCR was sufficiently low to minimise
329   production of chimeric artifacts, supporting previous findings (Laver et al., 2016). Several
330   methods have been developed for sequencing dsDNA viral metagenomes without skewing
331   relative abundance information important for comparative ecology, including an LASL approach
332   optimised for 454 sequencing (Duhaime et al., 2012; Hurwitz et al., 2013) and Nextera
333   sequencing (Roux et al., 2016b). Median per-genome coverages of VirION reads and short-read
334   Nextera datasets (5.6 M 2x300 bp paired-end) from the mock viral community were strongly
335   correlated (R^2=0.975, p<0.001, Figure 2a), indicating that the LASL approach used here for
336   multi-kilobasepair fragments retained relative abundance information observed in previous
337   LASL approaches.
338
339   Long-read and short-read assemblies of the mock viral community captured >99.7% of the six
340   mock viral community genomes (Table S1). Neither the short-read only assembly, hybrid
341   assembly nor long-read assemblies were able to capture all six genomes in six complete
342   contigs. Long-read methods gave the most contiguous assemblies, capturing the six genomes
343   across 14 contigs. In comparison, short-read only assemblies recovered the genomes across 26
344   contigs, whereas hybrid assembly reduced the number of contigs to 21. As expected, we
345   identified >250 times more indels errors in long-read only assemblies than in the short-read
346   assemblies scaffolded with long reads (average of 474 vs <2 indels per 100kbp, respectively).
347   Polishing of long-read only assemblies with short read data reduced the indel error rate to 22.78
348   per 100 kbp, indicating this was a successful strategy for reducing indel error of long-read
349   assemblies in metagenomic samples, but was not able to remove such errors completely. There
350   was no evidence of chimerism in any of the assemblies, indicating that Canu's *in silico*
351   correction of chimeras (Koren et al., 2017) successfully removed the low number of chimeric
352   sequences observed in the VirION reads during assembly.
353
354   *Combining VirION reads with short read data improves viral metagenomic assembly in an*
355   *environmental virome:* Long read sequencing of an environmental virome from the Western
356   English Channel produced 108,718 high quality VirION reads (median length: 3,625 bp; max
357   length: 17,019 bp, total yield of 0.39 Gbp). It is worth noting that recent developments of MinION
358   technology have improved flowcell yields to >10Gbp (*pers comms*). Therefore, our analyses
359   here represent low coverage of the viral community with long read data compared to currently
360   available (and fast-improving) technology.
361
362   Scaffolding short-read assemblies using VirION reads provided a small, but significant increase
363   in the number of putative viral genomes recovered (between 1.1 to 1.5-fold increase, Student t-
364   test, p<0.05) than short-read only assemblies up to a short-read sequencing depth of ~12Gbp
365   (Figure 3b). Above this depth, there was no significant difference between short-read
366   assemblies with and without scaffolding, suggesting assembly of short-read data was capturing
367   most of the viral community above this sequencing depth. For comparison, the median
368   sequencing depth of 137 Illumina sequenced viral metagenomes from the Global Ocean Virome

369   survey (study PRJEB4419 in the European Nucleotide Archive) was 8.67 Gbp (IQR=5.22 Gbp),
370   with 110 out of 137 samples sequenced to a depth of <12Gbp. Inclusion of VirION reads in
371   hybrid assemblies significantly increased the number of 'complete' (i.e. circular contigs) viral
372   genomes recovered once short-read sequencing depth increased above 12 Gbp (1.5 to 2.0-fold,
373   Student t-test, $p<0.05$) (Figure 3b). Details of differences in means and p-values at each depth
374   are available in Tables S3 and S4. When the full (30.8 Gbp) short-read dataset was used, the
375   inclusion of long reads for scaffolding De Bruijn Graph assemblies increased the median length
376   of recovered viral genomes by an average of 1.8 kbp compared to short-read only assemblies
377   (Mann-Whitney U test, n_1=1400, n_2=879, p-value<0.001). With an estimated mean gene density
378   of 1.4 genes per kb in phage dsDNA genomes (Mahmoudabadi & Phillips, 2018), this increased
379   length represents an extra 2.5 genes per contig.
380
381   Polishing of long-read assemblies of WEC VirION reads using complementary short-read data
382   removed a maximum of 172,854 insertion errors and 12,674 deletion errors (Figure S2). Error
383   correction reached an asymptote at ~9 Gbp of short-read sequencing data, with a median
384   coverage of ~70. As expected, the errors associated with long-read sequencing adversely
385   affected the lengths of protein predictions (Figure S3), in accordance with previous findings
386   (Warr & Watson, 2019). Proteins predicted from uncorrected VirION reads (median length of 72
387   aa, 70-74 aa 95%CI) were shorter (median difference = 61 aa, 69-53 aa 95%CI) than those
388   from RefSeq Caudovirales genomes (median length of 133 aa, 126-141aa 95%CI), and much
389   shorter (median difference = 88 aa, 83-95 95%CI) than those from the GOV dataset (median
390   length of 160 aa, 149-173 95%CI). Assembly of long reads with Canu includes a consensus-
391   based error-correction step (Koren et al., 2017), which increased median predicted protein
392   lengths to 87 aa (median difference of 15 aa, 14-15 95%CI) compared to raw VirION reads.
393   Polishing of long-read assemblies of VirION reads with short read data was highly effective in
394   restoring the length of predicted proteins (median length 127 aa, 120-135 aa 95%CI) to lengths
395   similar to those observed in RefSeq Caudovirales (median length = 133 aa, 126-141 aa 95%CI).
396   Proteins from polished reads had a median difference of -6 aa (-18-6 95%CI) compared to
397   RefSeq Caudovirales proteins. This suggests that not all frameshift errors were corrected in the
398   long-read assemblies, corroborated by evidence of increased indel errors observed in long-read
399   assemblies of mock viral community data compared to short-read assemblies.
400
401   Interestingly, predicted protein lengths from the GOV dataset (Roux et al., 2016a) (median
402   length = 160 aa), short-read only assembly of WEC virome (median length = 157 aa); hybrid
403   assembly of WEC virome (median length = 160 aa) and data from single-amplified viral
404   genomes (Martinez-Hernandez et al., 2017) (median length = 152 aa) were all of similar length
405   and 19 to 27 aa longer compared to those from RefSeq Caudovirales genomes, and 25 to 33 aa
406   longer than those from WEC polished long-read assemblies. In comparison, median predicted
407   protein length in 899 dsDNA phages was previously estimated at 136 aa (Mahmoudabadi &
408   Phillips, 2018) - similar to those found in our polished long-read assemblies from VirION reads.
409   Thus, either both the RefSeq Caudovirales dataset and that of Mahmoudabadi and Phillips are
410   under-representing longer viral predicted proteins found in marine viral metagenomes, or
411   predicted protein lengths in viral genomes from metagenomic data are longer than those
412   observed in cultured representatives. Whether this difference is biological or an artifact of

413   metagenomic assembly and gene calling is an interesting area for further investigation.
414
415   *Assembly and mapping of VirION reads captures more information about potential niche-*
416   *defining genomic islands than short-read only or hybrid assemblies*: In marine bacteria, genomic
417   islands have been identified as playing an important role in niche specialisation that drives
418   ecological speciation (Coleman et al., 2006). Genomic islands have also been found to be a
419   common feature of viral genomes and are typically enriched in functions associated with host
420   recognition (Mizuno, Ghai & Rodriguez-Valera, 2014). At all nucleotide identity cut-offs tested,
421   genomic islands captured on long-read assemblies were between 145 bp (112-184 bp 95%CI)
422   and 225 bp (189-259 bp, 95% CI) longer than those captured on short-read only or hybrid
423   assemblies. (Figure 4A, Figure S4A). There were no significant differences between the lengths
424   of genomic islands captured on short-read only or hybrid assemblies. The largest genomic
425   islands in each assembly type were 2.47 kbp, 5.75 kbp and 5.65 kbp in short-read only
426   assemblies, hybrid assemblies and long-read assemblies, respectively. In comparison, the
427   largest genomic islands identified in fosmid-based viral metagenomes were ~4.6 kbp (Mizuno,
428   Ghai & Rodriguez-Valera, 2014), suggesting that both hybrid and long-read approaches capture
429   similar length genomic islands as previous fosmid-based methods. Similarly, the density of GIs
430   was significantly greater in long-read assemblies (at between 40 bp (20-60 bp, 95%CI) and 100
431   bp (80-110 bp, 95%CI) of GI per kbp of genome) compared to short-read or hybrid assemblies
432   (Figure 4B, Figure S4B). Again, there was no significant difference between short-read only and
433   hybrid assemblies. At a nucleotide identity cut-off of 98% for read mapping, the length of GIs in
434   long-read assemblies were longer than those at 92% and 95%, (59 bp (18-106 bp, 95%CI) and
435   61 bp (13-105 bp, 95%CI) respectively), indicating that residual error in the polished reads may
436   be contributing to a slight increase in predicted GI length and density at high nucleotide identity.
437   However, these effect sizes are much smaller than those observed between long-read
438   assemblies and short and hybrid assemblies across all identity cut-offs, suggesting that long
439   reads do indeed improve the capture of genomic islands.
440
441   Previous work to identify viral genomic islands by recruiting short-reads back against assembled
442   contigs from fosmid libraries showed that nucleotide diversity within genomic islands was
443   associated with a constant-diversity model, with under-recruiting islands containing proteins
444   associated with host recognition and penetration, phage structure and DNA packaging structural
445   proteins (Mizuno, Ghai and Rodriguez-Valera, 2014). Here, we were able to gain further insight
446   into viral genomic islands by investigating whether diversity also occurred at the functional level.
447   VirION reads that spanned the full width of GIs were identified and their gene content was
448   predicted at the nucleotide level. Reads spanning the same GI were compared to see if different
449   proteins were encoded on different template strands prior to amplification.
450
451   In total, 137 genomic islands on 84 viral contigs had at least 10 VirION reads spanning their full
452   length. The 3,072 reads spanning these islands encoded 6,445 predicted proteins, of which
453   4,599 could be aligned to a protein within the NR database. Just 711 (15%) of aligned predicted
454   proteins returned a hit with known function, indicating that genomic islands are an important
455   source of 'genetic dark matter' (i.e. sequence of unknown function) in viral metagenomes
456   (Krishnamurthy & Wang 2017). In total 66 genomic islands contained genes with an assigned

457   function. These islands captured a range of functional proteins including those associated with
458   nucleotide biosynthesis; DNA methylation; redirection of host machinery; structural proteins and
459   associated chaperonins; endo and exonucleases and integrases (Table S5). 35 of the genomic
460   islands contained structural proteins (capsid, tail proteins, co-chaperonin GroES, YapH);
461   proteins associated with membrane recognition (carbohydrate-binding module, lectin-binding
462   proteins) or proteins associated with reconfiguring host metabolic machinery for viral synthesis
463   or defence suppression (RNA polymerase sigma factor, methyltransferases, tRNA synthetases,
464   anti-restriction protein), supporting the hypothesis that viral genomic islands are a hotspot for
465   Constant-Diversity evolutionary dynamics (Mizuno, Ghai and Rodriguez-Valera, 2014). Seven
466   out of 8 genomic islands containing a thymidylate synthase (an enzyme involved in pyrimidine
467   metabolism) also encoded partial hits to ribonucleotide reductase (involved in both purine and
468   pyrimidine metabolism). Ribonucleotide reductase has previously been identified as the
469   nucleotide metabolism gene most frequently interrupted by self-splicing introns (Dwivedi et al.,
470   2013). Similarly, thymidylate synthase has been found to contain self-splicing, group I introns in
471   phage genomes (Chu et al. 1984; Bechhofer, Hue and Shub, 1994), potentially identifying intron
472   splicing activity as a source of regulatory/functional variability and/or as a mechanism to
473   promote the movement of genetic material within the viral genomic islands in our data.
474   Functional, putatively niche-defining metabolic genes were also identified in the genomic
475   islands, including an ultraviolet light damage repair gene *uvsE* and genes associated with
476   photosystem II *(psbA)*. 25 out of the 66 genomic islands showed evidence of alternative gene
477   arrangements across their spanning reads, suggesting the content of genomic islands can vary
478   within viral populations at the structural, functional and nucleotide level.
479
480   *Assembly of VirION reads capture important, microdiverse populations previously missed by*
481   *short-read data*: It has been hypothesised that genomes assembled from short-read
482   metagenomes may be biased away from microdiverse populations (Martinez-Hernandez et al.,
483   2017; Roux et al., 2017). We reasoned that overlap layout consensus assembly of long reads,
484   followed by error correction might better capture genomes with high levels of microdiversity by
485   avoiding the unresolvable branches of De Bruijn Graph assemblies. We evaluated genome-level
486   nucleotide diversity (π) (Nei & Li, 1979) of both short-read assemblies and polished long-read
487   assemblies from the Western English Channel virome. Median levels of π were significantly (3-
488   fold) higher in polished long-read contigs than those derived from De Bruijn Graph assemblies
489   (two-sided Mann-Whitney U test: W=105,830, n_1= 758, n_2=206, p = 4.81 $\times$ 10^{-15}; Figure S5),
490   consistent with the hypothesis that long-read assembly of VirION reads captured genomes
491   previously lost due to failure to resolve assembly graphs as a consequence of microdiversity.
492
493   *Tig404 - an example of how VirION reads improve viral metagenomics:* An excellent example of
494   the benefit using VirION reads for viral metagenomics was found in a polished contig from long-
495   read assemblies that showed high nucleotide similarity and shared gene content to the globally
496   abundant pelagiphage HTVC010P (Zhao et al., 2013). This ecologically important virus and its
497   closely associated phages contain numerous genomic islands that comprise ~10% of their
498   genome and a shared 5.3 kbp genomic island containing a putative ribonuclease, bounded by
499   tail fibre proteins (Mizuno, Ghai & Rodriguez-Valera, 2014). It has also been predicted to
500   possess high microdiversity that challenges assembly from short-read data, leading to

501 fragmentation and thus under-representation in short-read viral metagenomes, but is
502 successfully captured using fosmid approaches and single-virus genomics (Mizuno et al., 2013;
503 Martinez-Hernandez et al., 2017). Clustering of viral contigs from the WEC by shared-gene
504 content using vContact2 (Bolduc et al., 2017) identified a virus called 'tig404' from long-read
505 assembly of VirION reads that was 89% identical at the nucleotide level to HTVC010P. We
506 mapped contigs from short-read only and hybrid assemblies against this genome at 95%
507 nucleotide identity over 80% of the length to evaluate the success of short-read and hybrid
508 assembly methods at capturing this genome, and identified its genomic islands as described
509 above (Figure 5). Both short-read only and hybrid assemblies were highly fragmented across
510 the genome. Analysis of median nucleotide diversity of tig404 was extremely high (Figure S5)
511 and provided supporting evidence that fragmentation may be a result of high microdiversity in
512 this phage. In contrast, VirION reads successfully overlapped across the genome and enabled
513 recovery of the genome through long-read assembly. Comparison of the genome of tig404 with
514 that of HTVC010P identified a shared genomic island containing a putative ribonuclease protein
515 and bounded by a tail fibre protein (Figure 5), similar to those observed in closely related taxa
516 from fosmid libraries (Mizuno, Ghai & Rodriguez-Valera, 2014).
517
518 In addition, we were able to exploit an additional benefit of long reads and use unpolished
519 VirION reads to explore the contents of the shared genomic island across the tig404 population
520 within the WEC virome. As each read is derived from a single DNA strand (excluding the low
521 abundance of chimeric reads), variance in the content of the genomic island within a population
522 would be captured on reads that align to the ends, or across, the genomic island. In total, 31
523 VirION reads extended from the boundaries into the genomic island (Figure 5). Of these, 17 had
524 sufficient overlap to use for identifying functional genes. Those at the 5' end of the genomic
525 island all contained a putative ribonuclease, whilst those at the 3' end all contained an internal
526 virion protein thought to be associated with puncturing the cell membrane in T7-like phages
527 (Mizuno, Ghai & Rodriguez-Valera, 2014). Thus, it would appear that, for this shared genomic
528 island at the population level, diversity occurs at the nucleotide level, rather than gene content
529 level. The fact that a similar gene content has now been found in the Western English Channel
530 (this study), the Sargasso Sea (Zhao et al., 2013) and the Mediterranean (Mizuno, Ghai &
531 Rodriguez-Valera, 2014) may indicate this is a conserved feature across the HTVC010P-like
532 phages. The encoding of a ribonuclease within a genomic island offers an interesting glimpse
533 into the host-virus interactions that occur during infection and suggests that degradation of RNA
534 is an important feature of the arms-race in HTVC010P-like phages with their *Pelagibacter* hosts.
535 Whether this is to shut down host metabolism, or to hijack host metabolism through
536 manipulation of regulatory machinery enriched in riboswitches (Meyer et al., 2009) requires
537 further investigation.
538
539 In total, analysis of VirION reads using a strategy to combine short and long read assemblies
540 (Figure 2) generated 2,645 putative viral contigs >10kbp from the Western English Channel. Of
541 these, 2,279 were from the de-replicated short and hybrid De Bruijn Graph assemblies and 366
542 from polished long-read assemblies. Our dataset represents the first virome sequenced from the
543 WEC and so we evaluated the global abundance of viral populations from the WEC by
544 competitive mapping of 10 million subsampled short reads from both the WEC and the GOV

545   dataset (Roux et al., 2016a). Representatives of viral populations from the WEC were then
546   pooled with those >10kb from the GOV dataset and other marine virome datasets (Table S2) to
547   make a total dataset of 20,545 viral contigs. Following competitive read recruitment with
548   FastViromeExplorer (Tithi et al., 2018), the top 50 most abundant viral genomes were identified
549   in each of the WEC and GOV surface samples. Out of 1,598 contigs, 81 of the most abundant
550   viral contigs were from long read assemblies of VirION reads from the WEC, representing a
551   significant enrichment (hypergeometric test for enrichment, p=6.6 x 10^{-19}). WEC contigs from
552   short-read only (42 contigs) and hybrid assemblies (77 contigs) were not significantly enriched
553   in the most abundant viral contigs. Thus, it is likely that long-read assembly of VirION reads
554   from the WEC captured important and globally abundant viral taxa previously missed in the
555   GOV datasets. Examination of relative abundance of WEC contigs in surface water samples
556   from the GOV showed that contigs from long-read assemblies of VirION reads recruited a large
557   proportion of the recruited reads from global samples, particularly in the Southern Atlantic
558   Ocean and waters off the Western coasts of Southern Africa and South America (Figure 6). In
559   total, clustering VirION-derived contigs from the Western English Channel with contigs from
560   previous studies (Table S2) by shared protein content produced 668 statistically supported viral
561   clusters. Of these, 202 contained contigs derived from long-read assembly of VirION reads, but
562   just 3 of these were comprised solely of these contigs. Thus, we are confident that previous
563   findings suggesting viral diversity at the genera level in surface oceans has been largely
564   documented (Roux et al., 2016a) are robust. Instead, we propose that long read assembly of
565   VirION reads provides greater phylogenetic resolution of viral clusters by capturing members
566   previously missed due to limitations in short-read assembly.
567

568   The most globally abundant and ubiquitous (identified in at least 10% of samples) viral genome
569   was a contig from a hybrid assembly, denoted H_NODE_1248 (Figure 7). This contig was 22.4
570   kbp in length and occupied a viral cluster (based on shared protein content) with 57 other
571   members, including vSAG-37-F6 (9th most abundant ubiquitous virus and 13th most abundant
572   across all samples), previously identified the most globally abundant virus (Martinez-Hernandez
573   et al., 2017, 2018). The viral cluster also contained 10 other contigs from long-read assembly of
574   VirION reads, ranging in size from 10 kbp to 27 kbp. Interestingly, pelagiphage HTVC010P,
575   once thought to be the most abundant virus on Earth (Zhao et al., 2013) was ranked 128th in
576   global abundance and did not meet the criteria of being both ubiquitous (identified in at least
577   10% of the samples) and abundant (in the top 100 most abundant viral taxa for each sample).
578   Upon its discovery as the most abundant global virus we previously urged a cautious
579   interpretation as any representative of a new viral clade will recruit reads from all similar viruses
580   in the environment (Zhao et al., 2013). As new representatives of these clades are captured in
581   metagenomic data it is likely that competitive recruitment of reads splits reads between all clade
582   members, reducing the estimated abundance of any one single member.
583

584   60% of the top 50 most abundant populations in the WEC were represented by a WEC contig
585   derived from long-read assemblies of VirION reads (Figure S7). The viral community in the
586   WEC sample was dominated by a 39,972 bp circular genome from a hybrid assembly. Denoted
587   H_NODE_525, this contig recruited 3.28 times more reads than the next most abundant contig
588   (Figure S7), but was not identified as globally abundant and ubiquitous (Figure 7). This virus

589 shared a viral cluster with the siphovirus *Pseudoalteromonas phage* vB_PspS-H6/1 but we were
590 not able to determine its putative host despite using a variety of tools (Ahlgren et al., 2016;
591 Galiez et al., 2017) (https://github.com/dutilh/CAT). A viral contig from hybrid assembly, denoted
592 H_NODE_6 was the longest complete viral genome identified in this study, with a 316 kbp
593 genome. In the short read-only assembly, this genome was broken into two contiguous contigs
594 of 204 kbp and 112 kbp, respectively (Figure S8). H_NODE_6 shared a viral cluster with the
595 myoviruses *Cronobacter sakasakii* phage GAP32 and Enterobacter phage vB_KleM-RaK2. At
596 359 kb and 346 kb respectively (Šimoliūnas et al., 2012; Abbasifar et al., 2014), these are some
597 of the largest phage genomes ever isolated. Recovery of this complete genome demonstrates
598 the capacity for hybrid assembly with VirION reads to capture complete genomes of very large
599 phages from complex communities on single contigs, which were fragmented using short-read
600 only assemblies.

601

602 **Conclusions**
603 In summary, this investigation represents the first use of long-read sequencing for viral
604 metagenomics. We have shown that using long-reads to scaffold short read De Bruijn Graph
605 assemblies improves recovery of complete viral genomes. Furthermore, overlap-layout
606 consensus assembly of VirION reads, followed by error correction with short reads captures
607 abundant and ubiquitous viral populations that are missed (possibly as a result of genome
608 fragmentation) by current short-read metagenomic methods. By combining these two
609 approaches, our proposed bioinformatics pipeline maximises the capture of viral diversity whilst
610 minimising the impact of high error rates associated with long-read sequencing and represents
611 a major addition to the viral metagenomics toolset. Improved capture of viral genomic islands
612 will enable better understanding of mechanisms underpinning host-virus interactions, as
613 demonstrated in our capture of a shared genomic island on the newly observed HTVC010P-like
614 pelagiphage tig404. Importantly, long-read sequencing on the MinION platform is undergoing
615 rapid improvements in terms of yield, with current technology providing at least an order of
616 magnitude more sequencing data than that produced in this study, at a cost of < $1000 per
617 flowcell. Thus, our approach represents a significant advantage in terms of cost, yield and
618 efficiency over fosmid and single-amplified genome approaches to capturing marine viruses that
619 are otherwise challenging to assemble.

620

621 As error rates associated with MinION technology continue to fall, we envisage less and less
622 complementary short-read data being required for polishing. A recent update to basecalling
623 methods has led to a significant reduction in indel errors and their associated impact on protein
624 prediction (Koren *et al.,* 2019). Furthermore, there is no technical reason to prevent our VirION
625 approach being used in conjunction with PacBio sequencing to further reduce error rates using
626 circular consensus sequencing. Such an approach would remove the need for short-read error
627 correction (Frank et al., 2016) and avoid the remaining indel errors observed following polishing
628 of MinION read assemblies with short-read data. Reductions in DNA input requirements and/or
629 improvements in DNA polymerases for increasing VirION amplicon lengths will further increase
630 its utility in recovering viral genomes from metagenomic samples.

631

632 Ultimately, community efforts to align the input requirements of long-read sequencing with DNA

633 recovery rates from viral communities will be rewarded by the ability to capture full-length viral
634 genomes on single reads (Houldcroft, Beale & Breuer, 2017), including all associated nucleotide
635 modifications (Viehweger et al., 2018). Oxford Nanopore sequencing interprets single stranded
636 nucleotides as they pass through the pore and there are significant efforts to develop protocols
637 for direct sequencing of RNA, dsDNA and ssDNA viruses (Keller et al., 2018; Viehweger et al.,
638 2018; McCabe et al., 2018). It is theoretically possible that with the right combination of ligases
639 and optimised buffers, we will soon be able to sequence dsDNA, ssDNA and RNA viruses, with
640 associated nucleotide modifications, within a single library preparation. We expect that the
641 VirION approach could be readily adapted for use with ssDNA or RNA viruses, provided
642 appropriate amplification of starting material could be achieved to meet the requirements for
643 sequencing. Therefore, the approach described here provides a significant step towards
644 capturing the full diversity of the viral community.
645
646 VirION offers a framework for robust downstream bioinformatic approaches to maximise the
647 benefits of long read sequencing, both now and as the technology continues to improve. Here,
648 we have shown that VirION long-read metagenomics of dsDNA viral communities offers the
649 potential to significantly improve our understanding of niche-differentiation, ecological speciation
650 and the role of viruses in microbial communities within aquatic (Roux et al., 2016a) and soil
651 (Pratama & van Elsas, 2018) environments, human health (Mirzaei & Maurice, 2017;
652 Aggarwala, Liang & Bushman, 2017) and industrial settings.
653

654 **Acknowledgements:** The authors thank the crew of the Plymouth Marine Laboratory vessel
655 'Quest' for collection of seawater samples, as well as Dr Simon Roux and Dr Benjamin Bolduc
656 for guidance and advice on bioinformatic analyses. Portions of this research were conducted
657 with high performance computing resources provided by Louisiana State University
658 (http://www.hpc.lsu.edu), Ohio Supercomputer Center (Center, 1987), and the HPC
659 infrastructure at University of Exeter.

660
661

References

663

664 Abbasifar R., Griffiths MW., Sabour PM., Ackermann H-W., Vandersteegen K., Lavigne R.,
665   Noben J-P., Alanis Villa A., Abbasifar A., Nash JHE., Kropinski AM. 2014. Supersize me:
666   *Cronobacter sakazakii* phage GAP32. *Virology* 460-461:138–146.
667 Aggarwala V., Liang G., Bushman FD. 2017. Viral communities of the human gut: metagenomic
668   analysis of composition and dynamics. *Mobile DNA* 8:12.
669 Ahlgren NA., Ren J., Lu YY., Fuhrman JA., Sun F. 2016. Alignment-free oligonucleotide
670   frequency dissimilarity measure improves prediction of hosts from metagenomically-derived
671   viral sequences. *Nucleic acids research* 45:39–53.
672 Ashton PM., Nair S., Dallman T., Rubino S., Rabsch W., Mwaigwisya S., Wain J., O'Grady J.
673   2015. MinION nanopore sequencing identifies the position and structure of a bacterial

674   antibiotic resistance island. *Nature biotechnology* 33:296–300.

675 Bechhofer DH, Hue KK, Shub DA. 1994. An intron in the thymidylate synthase gene of Bacillus
676   bacteriophage beta 22: evidence for independent evolution of a gene, its group I intron, and
677   the intron open reading frame. *Proceedings of the National Academy of Sciences of the*
678   *United States of America* 91:11669–11673.

679 Bolduc B., Jang HB., Doulcier G., You Z-Q., Roux S., Sullivan MB. 2017. vConTACT: an iVirus
680   tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ*
681   5:e3243.

682

683 Bray NL., Pimentel H., Melsted P., Pachter L. 2016. Near-optimal probabilistic RNA-seq
684   quantification. *Nature biotechnology* 34:525–527.

685 Breitbart M., Thompson LR., Suttle CA., Sullivan MB. 2007. Exploring the Vast Diversity of
686   Marine Viruses. *Oceanography* 20:135–139.

687 Brocchieri L., Karlin S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic*
688   *acids research* 33:3390–3400.

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

695 Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. 2009.
696   BLAST+: architecture and applications. *BMC bioinformatics* 10:421.

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

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

701 Coleman ML., Sullivan MB., Martiny AC., Steglich C., Barry K., Delong EF., Chisholm SW.
702   2006. Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science*
703   311:1768–1770.

704 Cumming G. 2014. The new statistics: why and how. *Psychological science* 25:7–29.

705 Cunningham BR., Brum JR., Schwenck SM., Sullivan MB., John SG. 2015. An inexpensive,
706   accurate, and precise wet-mount method for enumerating aquatic viruses. *Applied and*
707   *environmental microbiology* 81:2995–3000.

708 Dabney A., Storey JD., Warnes GR. 2015. qvalue: Q-value estimation for false discovery rate
709   control. *R package version 2.12.0*.

710 Darling ACE., Mau B., Blattner FR., Perna NT. 2004. Mauve: multiple alignment of conserved
711   genomic sequence with rearrangements. *Genome research* 14:1394–1403.

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

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

717 Driscoll CB., Otten TG., Brown NM., Dreher TW. 2017. Towards long-read metagenomics:

718   complete assembly of three novel genomes from bacteria dependent on a diazotrophic
719   cyanobacterium in a freshwater lake co-culture. *Standards in genomic sciences* 12:9.
720 Duhaime MB., Deng L., Poulos BT., Sullivan MB. 2012. Towards quantitative metagenomics of
721   wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and
722   optimization of the linker amplification method. *Environmental microbiology* 14:2526–2537.
723 Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M. 2013. A bioinformatic analysis of
724   ribonucleotide reductase genes in phage genomes and metagenomes. *BMC evolutionary*
725   *biology* 13:33.
726 Forterre P. 2013. The virocell concept and environmental microbiology. *The ISME journal*
727   7:233–236.
728 Frank JA., Pan Y., Tooming-Klunderud A., Eijsink VGH., McHardy AC., Nederbragt AJ., Pope
729   PB. 2016. Improved metagenome assemblies and taxonomic binning using long-read
730   circular consensus sequence data. *Scientific reports* 6:25373.
731 Galiez C., Siebert M., Enault F., Vincent J., Söding J. 2017. WIsH: who is the host? Predicting
732   prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33:3113–3114.
733 Giovannoni S. 2017. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annual*
734   *review of marine science* 9:231–255.
735 Houldcroft CJ, Beale MA, Breuer J. 2017. Clinical and biological insights from viral genome
736   sequencing. *Nature reviews. Microbiology* 15:183–192.
737 Hurwitz BL., Deng L., Poulos BT., Sullivan MB. 2013. Evaluation of methods to concentrate and
738   purify ocean virus communities through comparative, replicated metagenomics.
739   *Environmental microbiology* 15:1428–1440.
740 Hurwitz BL., Hallam SJ., Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit
741   and dark ocean. *Genome biology* 14:R123.
742 Hurwitz BL., Sullivan MB. 2013. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic
743   Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PloS one* 8:e57355.
744 Hurwitz BL., U'Ren JM. 2016. Viral metabolic reprogramming in marine ecosystems. *Current*
745   *opinion in microbiology* 31:161–168.
746 Jain M., Fiddes IT., Miga KH., Olsen HE., Paten B., Akeson M. 2015. Improved data analysis for
747   the MinION nanopore sequencer. *Nature methods* 12:351–356.
748 Jain M., Koren S., Miga KH., Quick J., Rand AC., Sasani TA., Tyson JR., Beggs AD., Dilthey
749   AT., Fiddes IT., Malla S., Marriott H., Nieto T., O'Grady J., Olsen HE., Pedersen BS., Rhie
750   A., Richardson H., Quinlan AR., Snutch TP., Tee L., Paten B., Phillippy AM., Simpson JT.,
751   Loman NJ., Loose M. 2018. Nanopore sequencing and assembly of a human genome with
752   ultra-long reads. *Nature biotechnology* 36:338–345.
753 John SG., Mendez CB., Deng L., Poulos B., Kauffman AK., Kern S., Brum J., Polz MF., Boyle
754   EA., Sullivan MB. 2011. A simple and efficient method for concentration of ocean viruses by
755   chemical flocculation. *Environmental microbiology reports* 3:195–202.
756 Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A.,
757   Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012. Geneious
758   Basic: an integrated and extendable desktop software platform for the organization and
759   analysis of sequence data. *Bioinformatics* 28:1647–1649.
760 Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ, Neuhaus EB,
761   Dugan VG, Wentworth DE, Barnes JR. 2018. Direct RNA Sequencing of the Coding

762   Complete Influenza A Virus Genome. *Scientific reports* 8:14408.

763   Koren S., Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from
764   long-read sequencing and assembly. *Current opinion in microbiology* 23:110–120.

765   Koren S., Walenz BP., Berlin K., Miller JR., Bergman NH., Phillippy AM. 2017. Canu: scalable
766   and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
767   *Genome research* 27:722–736.

768   Koren S, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2019. Reply to "Errors in long-read
769   assemblies can critically affect protein prediction." *Nature biotechnology*:1.
770   doi:10.1038/s41587-018-0005-y

771   Krishnamurthy SR, Wang D. 2017. Origins and challenges of viral dark matter. *Virus research*
772   239:136–142.

773   Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*
774   9:357–359.

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

778   Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin
779   R., 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map
780   format and SAMtools. *Bioinformatics* 25:2078–2079.

781   Loman NJ., Quick J., Simpson JT. 2015. A complete bacterial genome assembled de novo
782   using only nanopore sequencing data. *Nature methods* 12:733–735.

783   Luo E., Aylward FO., Mende DR., DeLong EF. 2017. Bacteriophage Distributions and Temporal
784   Variability in the Ocean's Interior. *mBio* 8. DOI: 10.1128/mBio.01903-17.

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

787   Martinez-Hernandez F., Fornas O., Lluesma Gomez M., Bolduc B., de la Cruz Peña MJ.,
788   Martínez JM., Anton J., Gasol JM., Rosselli R., Rodriguez-Valera F., Sullivan MB., Acinas
789   SG., Martinez-Garcia M. 2017. Single-virus genomics reveals hidden cosmopolitan and
790   abundant viruses. *Nature communications* 8:15892.

791   Martinez-Hernandez F., Fornas Ò., Lluesma Gomez M., Garcia-Heredia I., Maestre-Carballa L.,
792   López-Pérez M., Haro-Moreno JM., Rodriguez-Valera F., Martinez-Garcia M. 2018. Single-
793   cell genomics uncover *Pelagibacter* as the putative host of the extremely abundant
794   uncultured 37-F6 viral population in the ocean. *The ISME journal*. DOI: 10.1038/s41396-
795   018-0278-7.

796   McCabe M, Cormican P, Johnston D, Earley B. 2018. Simultaneous detection of DNA and RNA
797   virus species involved in bovine respiratory disease by PCR-free rapid tagmentation-based
798   library preparation and MinION nanopore sequencing. *bioRxiv*:269936. DOI:
799   10.1101/269936.

800   Merchant N., Lyons E., Goff S., Vaughn M., Ware D., Micklos D., Antin P. 2016. The iPlant
801   Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences.
802   *PLoS biology* 14:e1002342.

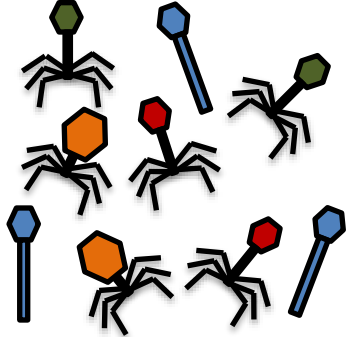
803   Meyer MM., Ames TD., Smith DP., Weinberg Z., Schwalbach MS., Giovannoni SJ., Breaker RR.
804   2009. Identification of candidate structured RNAs in the marine organism "*Candidatus*
805   Pelagibacter ubique." *BMC genomics* 10:268.

806 Mikheenko A., Saveliev V., Gurevich A. 2016. MetaQUAST: evaluation of metagenome
807   assemblies. *Bioinformatics* 32:1088–1090.
808 Mirzaei MK., Maurice CF. 2017. Ménage à trois in the human gut: interactions between host,
809   bacteria and phages. *Nature reviews. Microbiology*. DOI: 10.1038/nrmicro.2017.30.
810 Mizuno CM., Ghai R., Rodriguez-Valera F. 2014. Evidence for metaviromic islands in marine
811   phages. *Frontiers in microbiology* 5. DOI: 10.3389/fmicb.2014.00027.
812 Mizuno CM., Ghai R., Saghaï A., López-García P., Rodriguez-Valera F. 2016. Genomes of
813   Abundant and Widespread Viruses from the Deep Ocean. *mBio* 7. DOI:
814   10.1128/mBio.00805-16.
815 Mizuno CM., Rodriguez-Valera F., Kimes NE., Ghai R. 2013. Expanding the marine virosphere
816   using metagenomics. *PLoS genetics* 9:e1003987.
817 Nei M., Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction
818   endonucleases. *Proceedings of the National Academy of Sciences of the United States of
819   America* 76:5269–5273.
820 Nepusz T., Yu H., Paccanaro A. 2012. Detecting overlapping protein complexes in protein-
821   protein interaction networks. *Nature methods* 9:471–472.
822 Noble RT. 2001. Enumeration of viruses. In: *Methods in Microbiology*. Academic Press, 43–51.
823 Noguchi H., Taniguchi T., Itoh T. 2008. MetaGeneAnnotator: detecting species-specific patterns
824   of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage
825   genomes. *DNA research: an international journal for rapid publication of reports on genes
826   and genomes* 15:387–396.
827 Nurk S., Meleshko D., Korobeynikov A., Pevzner PA. 2017. metaSPAdes: a new versatile
828   metagenomic assembler. *Genome research* 27:824–834.
829 Olson ND., Treangen TJ., Hill CM., Cepeda-Espinoza V., Ghurye J., Koren S., Pop M. 2017.
830   Metagenomic assembly through the lens of validation: recent advances in assessing and
831   improving the quality of genomes assembled from metagenomes. *Briefings in
832   bioinformatics*. DOI: 10.1093/bib/bbx098.
833 Pratama AA., van Elsas JD. 2018. The "Neglected" Soil Virome - Potential Role and Impact.
834   *Trends in microbiology* 26:649–662.
835 Roux S., Brum JR., Dutilh BE., Sunagawa S., Duhaime MB., Loy A., Poulos BT., Solonenko N.,
836   Lara E., Poulain J., Pesant S., Kandels-Lewis S., Dimier C., Picheral M., Searson S.,
837   Cruaud C., Alberti A., Duarte CM., Gasol JM., Vaque D., Tara Oceans C., Bork P., Acinas
838   SG., Wincker P., Sullivan MB. 2016a. Ecogenomics and potential biogeochemical impacts
839   of globally abundant ocean viruses. *Nature* 537:689–693.
840 Roux S., Emerson JB., Eloe-Fadrosh EA., Sullivan MB. 2017. Benchmarking viromics: An in
841   silico evaluation of metagenome-enabled estimates of viral community composition and
842   diversity. *PeerJ*. DOI: 10.7717/peerj.3817.
843 Roux S., Enault F., Hurwitz BL., Sullivan MB. 2015. VirSorter: mining viral signal from microbial
844   genomic data. *PeerJ* 3:e985.
845 Roux S., Solonenko NE., Dang VT., Poulos BT., Schwenck SM., Goldsmith DB., Coleman ML.,
846   Breitbart M., Sullivan MB. 2016b. Towards quantitative viromics for both double-stranded
847   and single-stranded DNA viruses. *PeerJ* 4:e2777.
848 Shagin DA., Lukyanov KA., Vagner LL., Matz MV. 1999. Regulation of average length of
849   complex PCR product. *Nucleic acids research* 27:e23.

850 Šimoliūnas E., Kaliniene L., Truncaite L., Klausa V., Zajančkauskaite A., Meškys R. 2012.
851   Genome of *Klebsiella sp.*-infecting bacteriophage vB_KleM_RaK2. *Journal of virology*
852   86:5406.
853 Sullivan MB. 2015. Viromes, not gene markers, for studying double-stranded DNA virus
854   communities. *Journal of virology* 89:2459–2461.
855 Suttle CA. 2005. Viruses in the sea. *Nature* 437:356–361.
856 Suttle CA. 2007. Marine viruses — major players in the global ecosystem. *Nature reviews.*
857   *Microbiology* 5:801–812.
858 Temperton B., Giovannoni SJ. 2012. Metagenomics: microbial diversity through a scratched
859   lens. *Current opinion in microbiology* 15:605–612.
860 Thompson LR., Sanders JG., McDonald D., Amir A., Ladau J., Locey KJ., Prill RJ., Tripathi A.,
861   Gibbons SM., Ackermann G., Navas-Molina JA., Janssen S., Kopylova E., Vázquez-Baeza
862   Y., González A., Morton JT., Mirarab S., Zech Xu Z., Jiang L., Haroon MF., Kanbar J., Zhu
863   Q., Jin Song S., Kosciolek T., Bokulich NA., Lefler J., Brislawn CJ., Humphrey G., Owens
864   SM., Hampton-Marcell J., Berg-Lyons D., McKenzie V., Fierer N., Fuhrman JA., Clauset A.,
865   Stevens RL., Shade A., Pollard KS., Goodwin KD., Jansson JK., Gilbert JA., Knight R.,
866   Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's
867   multiscale microbial diversity. *Nature* 551:457–463.
868 Tithi SS., Aylward FO., Jensen RV., Zhang L. 2018. FastViromeExplorer: a pipeline for virus
869   and phage identification and abundance profiling in metagenomics data. *PeerJ* 6:e4227.
870 Torsvik V., Øvreås L. 2002. Microbial diversity and function in soil: from genes to ecosystems.
871   *Current opinion in microbiology* 5:240–245.
872 Treusch AH., Vergin KL., Finlay LA., Donatz MG., Burton RM., Carlson CA., Giovannoni SJ.
873   2009. Seasonality and vertical structure of microbial communities in an ocean gyre. *The*
874   *ISME journal* 3:1148–1163.
875 Viehweger A, Krautwurst S, Lamkiewicz K, Madhugiri R, Ziebuhr J, Hölzer M, Marz M. 2018.
876   Nanopore direct RNA sequencing reveals modification in full-length coronavirus genomes.
877   *bioRxiv*:483693. DOI: 10.1101/483693.
878 Warr A. & Watson M. (2019). Errors in long-read assemblies can critically affect protein
879   prediction. Nature Biotechnology, 1. http://dx.doi.org/10.1038/s41587-018-0004-z
880 Walker BJ., Abeel T., Shea T., Priest M., Abouelliel A., Sakthikumar S., Cuomo CA., Zeng Q.,
881   Wortman J., Young SK., Earl AM. 2014. Pilon: an integrated tool for comprehensive
882   microbial variant detection and genome assembly improvement. *PloS one* 9:e112963.
883
884 Weinbauer MG. 2004. Ecology of prokaryotic viruses. *FEMS microbiology reviews* 28:127–181.
885 Weirather JL., de Cesare M., Wang Y., Piazza P., Sebastiano V., Wang X-J., Buck D., Au KF.
886   2017. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore
887   Technologies and their applications to transcriptome analysis. *F1000Research* 6:100.
888 Wick RR., Judd LM., Gorrie CL., Holt KE. 2017. Unicycler: Resolving bacterial genome
889   assemblies from short and long sequencing reads. *PLoS computational biology*
890   13:e1005595.
891 Zhao Y., Temperton B., Thrash JC., Schwalbach MS., Vergin KL., Landry ZC., Ellisman M.,
892   Deerinck T., Sullivan MB., Giovannoni SJ. 2013. Abundant SAR11 viruses in the ocean.
893   *Nature* 494:357–360.

894

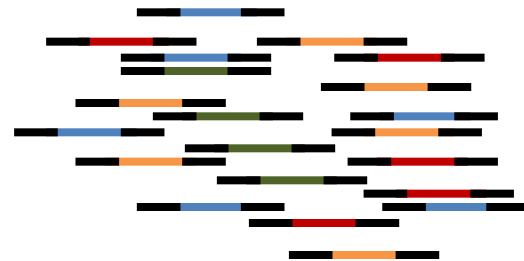
Workflow for preparation of free-viral fraction DNA for MinION sequencing

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.

FFR to concentrate free viral particles

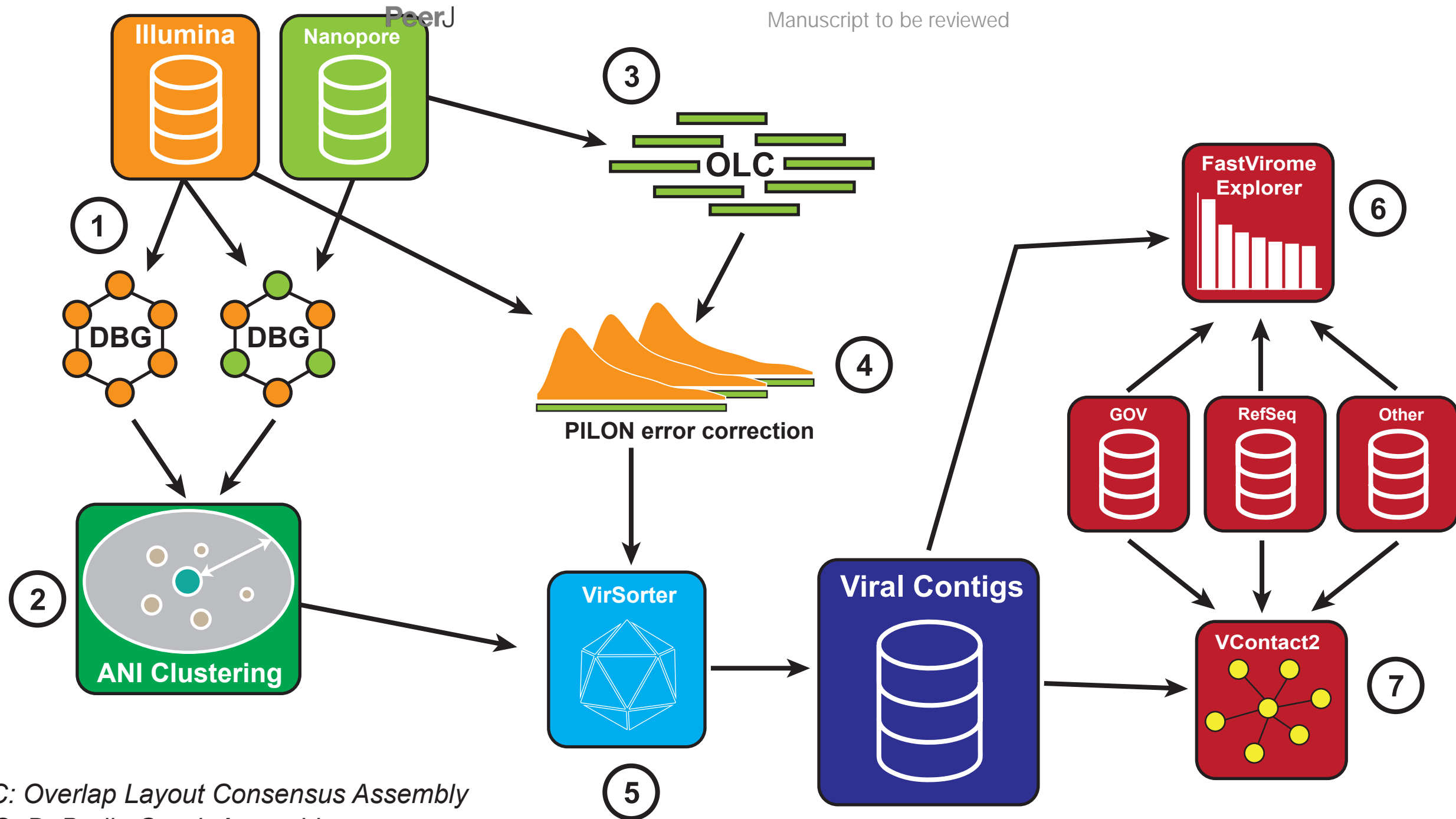
DNA extraction and clean-up; quantification: 50-100 ng

Shearing to 8-9 kbp using a g-TUBE (Covaris)

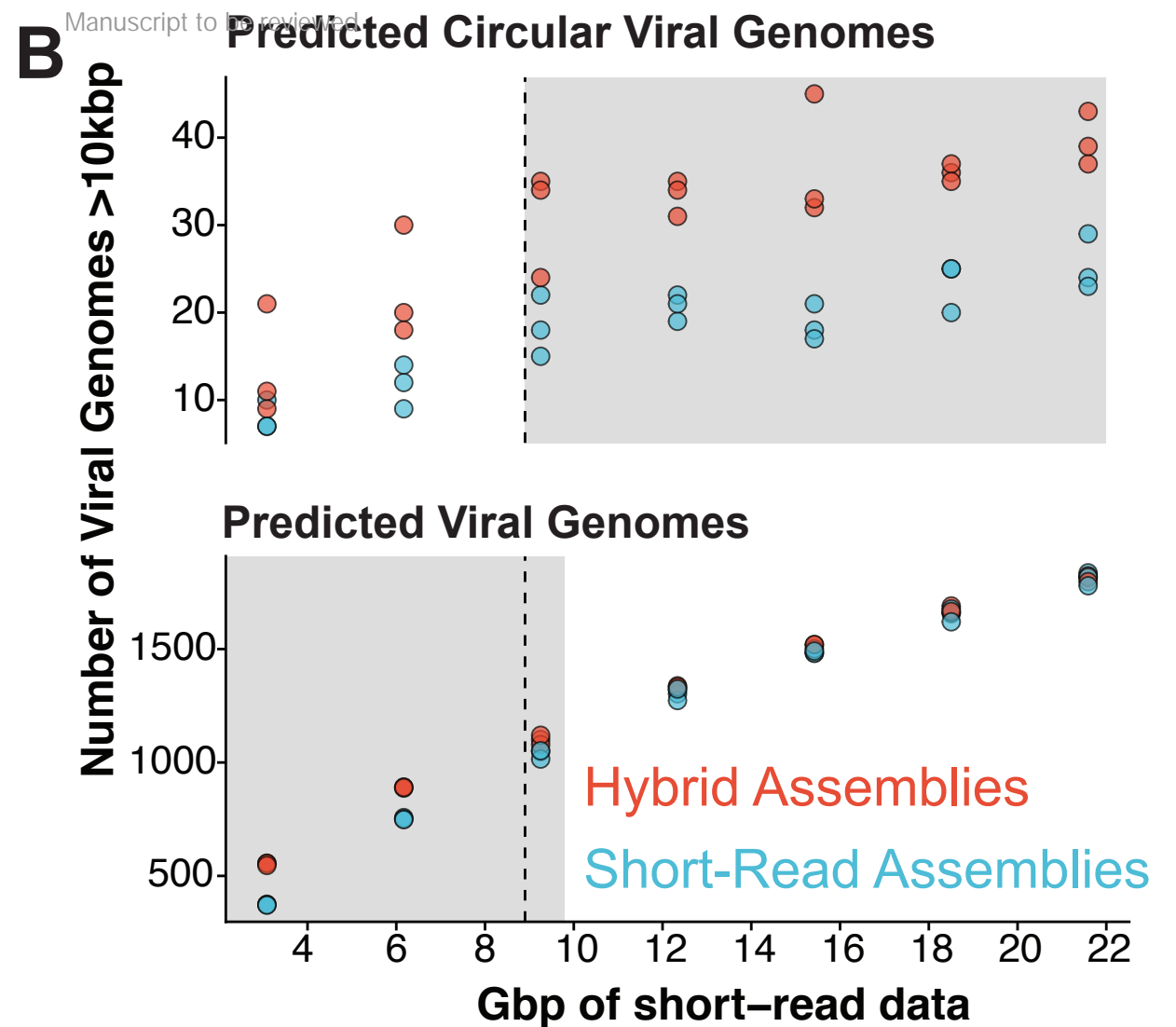
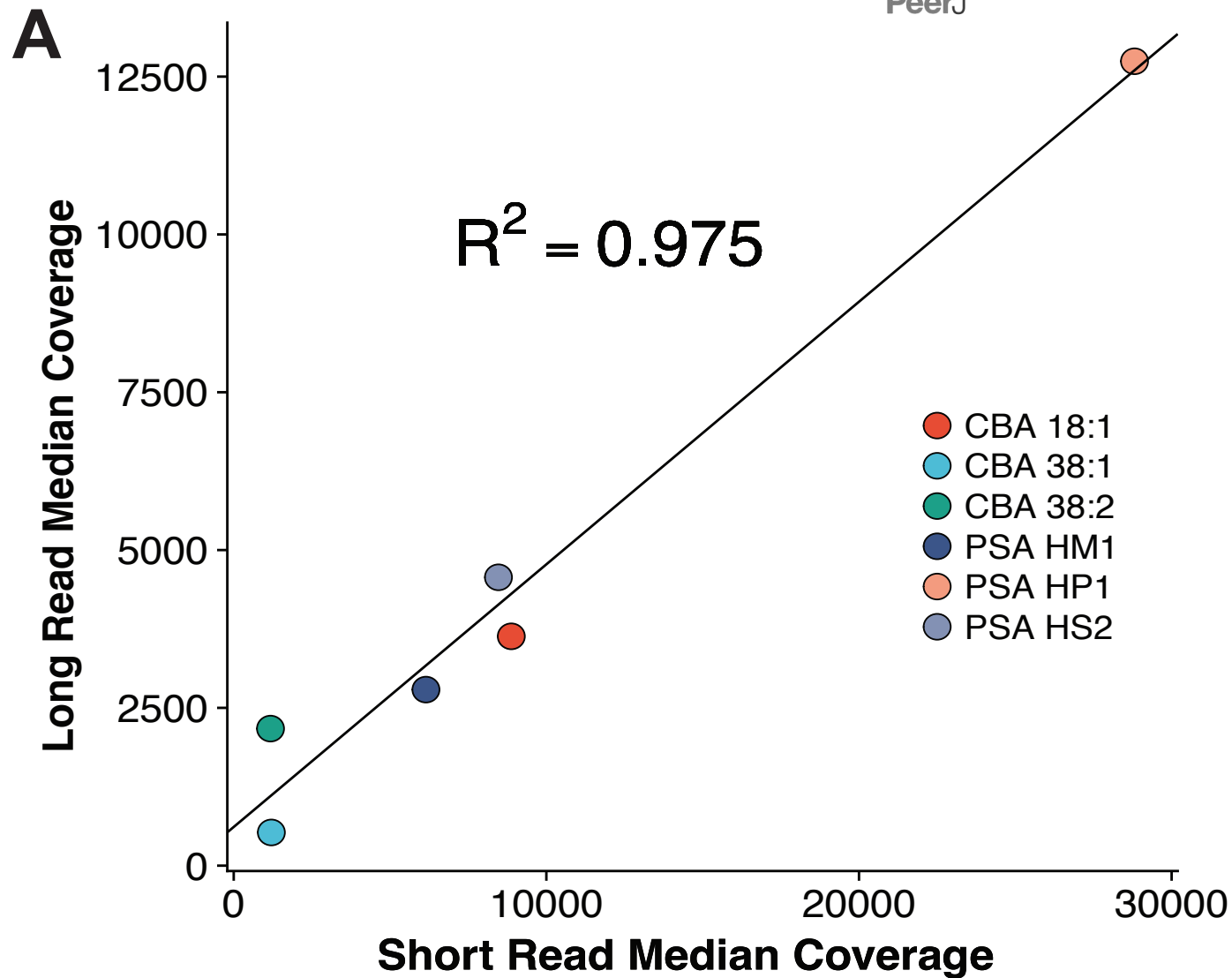
MinION Library Prep (adapter ligation)

15 Cycles of PCR amplification to yield ~1µg of viral amplicon DNA

PCR adapter ligation

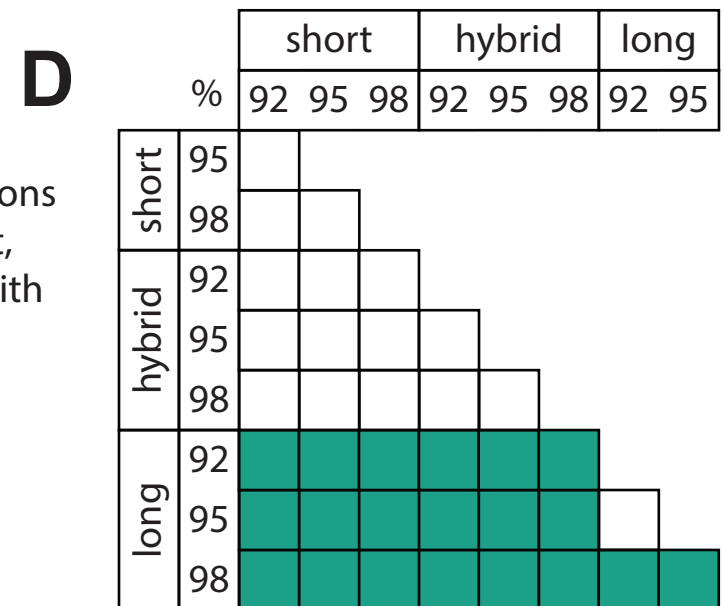
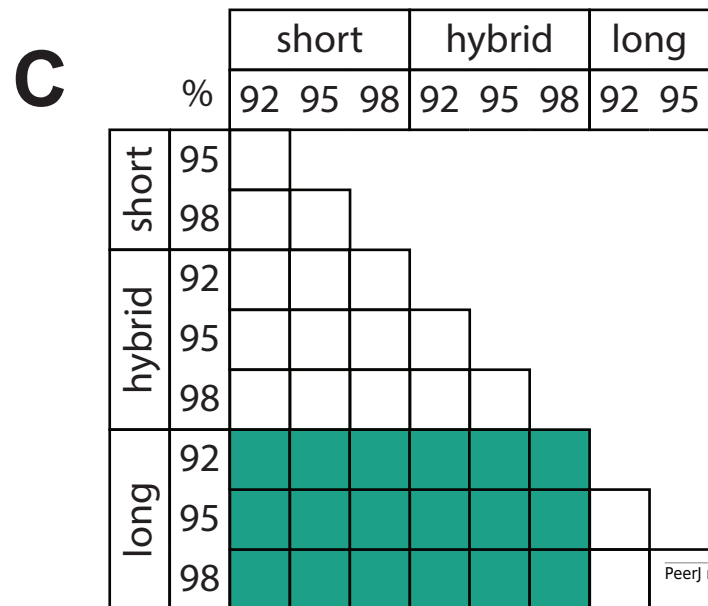
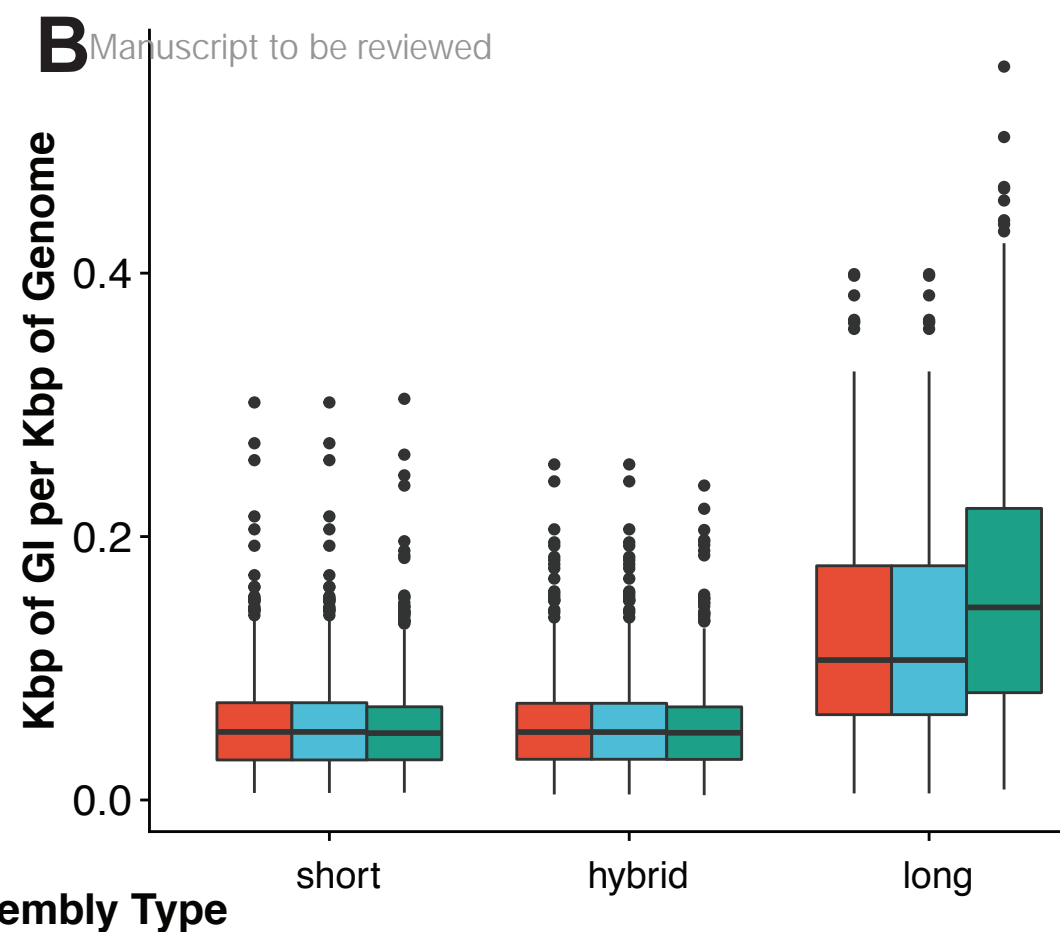
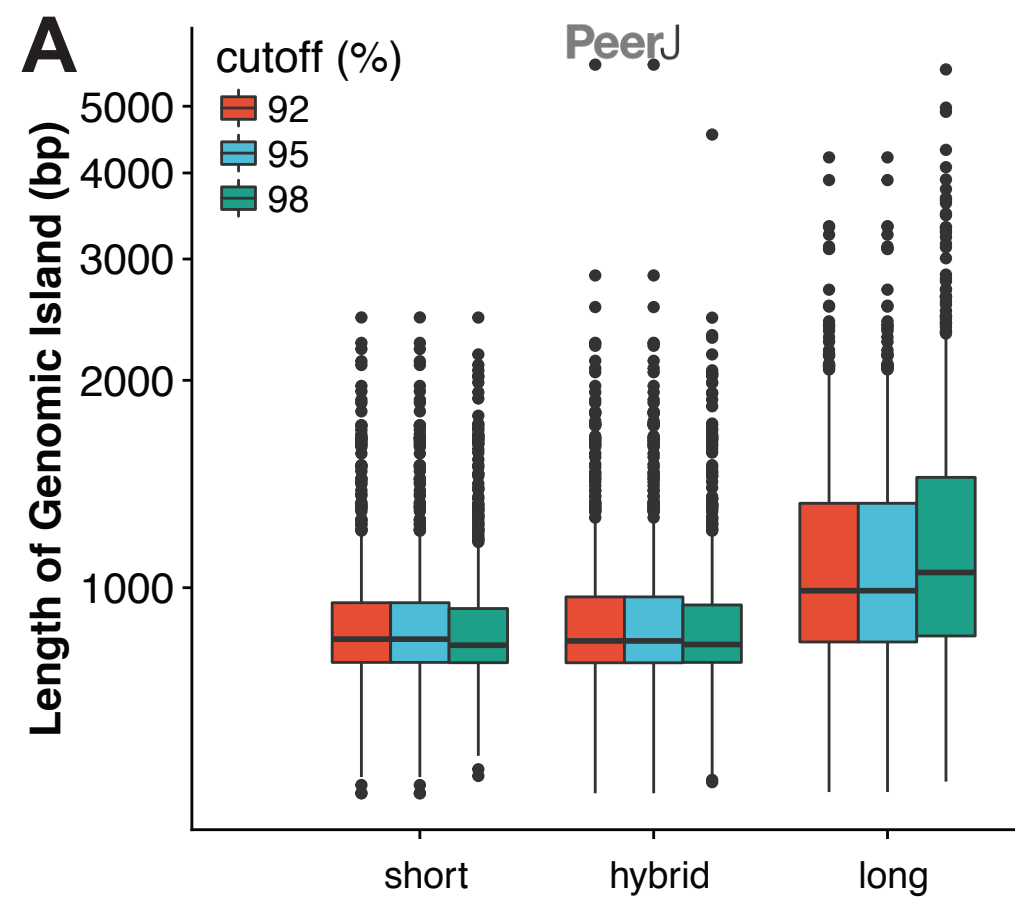
Bioinformatics pipeline for VirION reads and complementary short-read sequencing for viral metagenomes

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 90% 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 Supplementary Table 2) using vConTACT2 (Bolduc et al., 2017).

OLC: Overlap Layout Consensus Assembly
DBG: DeBrujin Graph Assembly
ANI: Average Nucleotide Identity
GOV: Global Ocean Virome

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).

Dashed line represents mean size of Global Ocean Virome Datasets

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 Figure S4.

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.

A coverage

B π_{max}

C

Long read OLC Assembly

Predicted Genomic Island

Hybrid DBG Assembly

Short Read DBG Assembly

D

VirION reads > 2500 bp

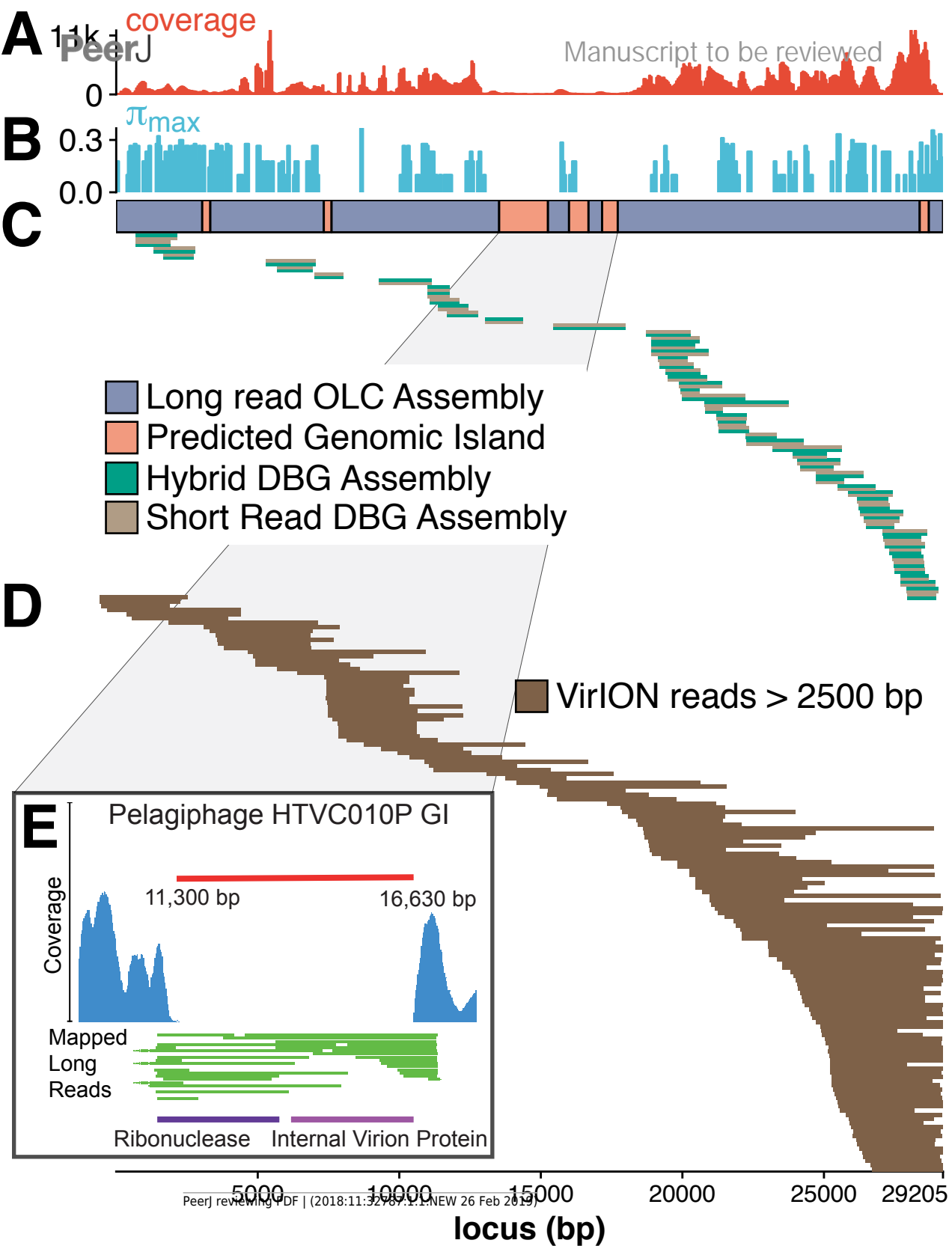
E Pelagiphage HTVC010P GI

11,300 bp   16,630 bp

Coverage

Mapped
Long
Reads

Ribonuclease   Internal Virion Protein

locus (bp)

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) ; '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, 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 '*'.

WEC VirION hybrid

WEC VirION short

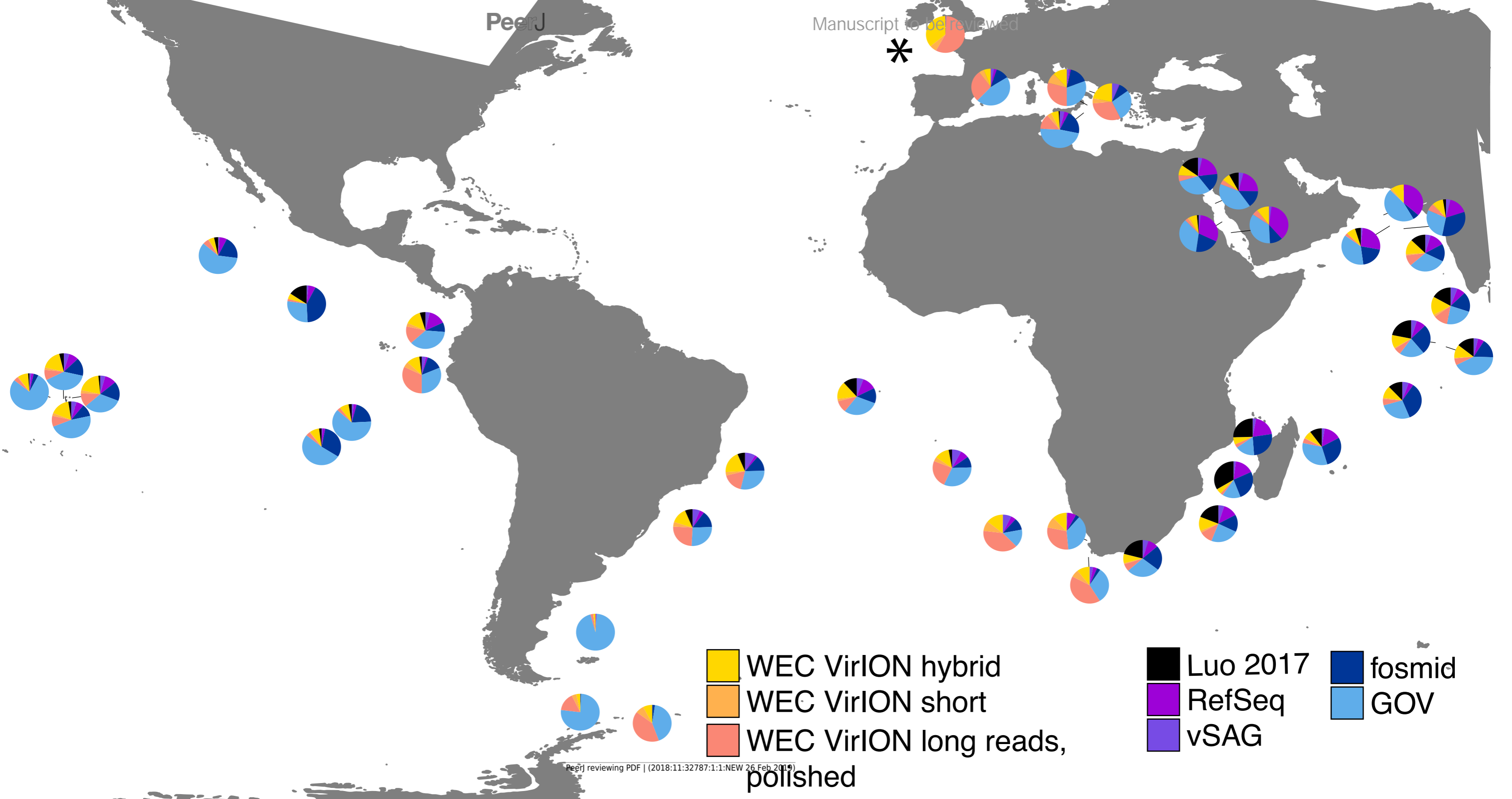
WEC VirION long reads,
polished

Luo 2017

RefSeq

vSAG

fosmid

GOV

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 (Supplementary Table 2); 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-maximization 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