VirION2: a short- and long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature

Olivier Zablocki [1,2], **Michelle Michelsen** [3], **Marie Burris** [1], **Natalie Solonenko** [1], **Joanna Warwick-Dugdale** [3,4], **Romik Ghosh** [1], **Jennifer Pett-Ridge** [5], **Matthew B Sullivan** [1,2,6], **Ben Temperton** [Corresp. 3]

[1] Department of Microbiology, The Ohio State University, Columbus, OH, United States

[2] Center of Microbiome Science, The Ohio State University, Columbus, OH, United States

[3] School of Biosciences, University of Exeter, Exeter, Devon, United Kingdom

[4] Plymouth Marine Laboratory, Plymouth, Devon, United Kingdom

[5] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, United States

[6] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, United States

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

Microbes play fundamental roles in shaping natural ecosystem properties and functions, but do so under constraints imposed by their viral predators. However, studying viruses in nature can be challenging due to low biomass and the lack of universal gene markers. Though metagenomic short-read sequencing has greatly improved our virus ecology toolkit— and revealed many critical ecosystem roles for viruses — microdiverse populations and fine-scale genomic traits are missed. Some of these microdiverse populations are abundant and the missed regions may be of interest for identifying selection pressures that underpin evolutionary constraints associated with hosts and environments. Though long-read sequencing promises complete virus genomes on single reads, it currently suffers from high DNA requirements and sequencing errors that limit accurate gene prediction. Here we introduce VirION2, an integrated short- and long-read metagenomic wet-lab and informatics pipeline that updates our previous method (VirION) to further enhance the utility of long-read viral metagenomics. Using a viral mock community, we first optimized laboratory protocols (polymerase choice, DNA shearing size, PCR cycling) to enable 76% longer reads (now median length of 6,965 bp) from 100-fold less input DNA (now 1 nanogram). Using a virome from a natural seawater sample, we compared viromes generated with VirION2 against other library preparation options (unamplified, original VirION, and short-read), and optimized downstream informatics for improved long-read error correction and assembly. VirION2 assemblies combined with short-read based data ('enhanced' viromes), provided significant improvements over VirION libraries in the recovery of longer and more complete viral genomes, and our optimized error-correction strategy using long- and short-read data achieved 99.97%

accuracy. In the seawater virome, VirION2 assemblies captured 5,161 viral populations (including all of the virus populations observed in the other assemblies), 30% of which were uniquely assembled through inclusion of long-reads, and 22% of the top 10% most abundant virus populations derived from assembly of long-reads. Viral populations unique to VirION2 assemblies had significantly higher microdiversity means, which may explain why short-read virome approaches failed to capture them. These findings suggest the VirION2 sample prep and workflow can help researchers better investigate the virosphere, even from challenging low-biomass samples. Our new protocols are available to the research community on protocols.io as a 'living document' to facilitate dissemination of updates to keep pace with the rapid evolution of long-read sequencing technology.

1 **VirION2: a short- and long-read sequencing and informatics workflow to study the**
2 **genomic diversity of viruses in nature**

3 Olivier Zablocki[1,2], Michelle Michelsen[3], Marie Burris[1], Natalie Solonenko[1], Joanna Warwick-Dugdale[3,4],
4 Romik Ghosh[1], Jennifer Pett-Ridge[5], Matthew B. Sullivan[1,2,6], and Ben Temperton[3]

5 [1] Department of Microbiology, Ohio State University, Columbus, OH 43210, United States
6 [2] Center of Microbiome Science, Ohio State University, Columbus, OH 43210, United States
7 [3] School of Biosciences, University of Exeter, Stocker Road, Exeter, EX4 4QD, United Kingdom
8 [4] Plymouth Marine Laboratory, Prospect Place, The Hoe, Plymouth PL1 3DH, United Kingdom
9 [5] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA
10   94550, United States
11 [6]Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH
12 43210, United States
13
14 * Corresponding Author:
15 Ben Temperton
16 b.temperton@exeter.ac.uk

18 **ABSTRACT**

19 Microbes play fundamental roles in shaping natural ecosystem properties and functions, but do so under
20 constraints imposed by their viral predators. However, studying viruses in nature can be challenging due
21 to low biomass and the lack of universal gene markers. Though metagenomic short-read sequencing has
22 greatly improved our virus ecology toolkit— and revealed many critical ecosystem roles for viruses —
23 microdiverse populations and fine-scale genomic traits are missed. Some of these microdiverse
24 populations are abundant and the missed regions may be of interest for identifying selection pressures that
25 underpin evolutionary constraints associated with hosts and environments. Though long-read sequencing
26 promises complete virus genomes on single reads, it currently suffers from high DNA requirements and
27 sequencing errors that limit accurate gene prediction. Here we introduce VirION2, an integrated short-
28 and long-read metagenomic wet-lab and informatics pipeline that updates our previous method (VirION)
29 to further enhance the utility of long-read viral metagenomics. Using a viral mock community, we first
30 optimized laboratory protocols (polymerase choice, DNA shearing size, PCR cycling) to enable 76%
31 longer reads (now median length of 6,965 bp) from 100-fold less input DNA (now 1 nanogram). Using a
32 virome from a natural seawater sample, we compared viromes generated with VirION2 against other
33 library preparation options (unamplified, original VirION, and short-read), and optimized downstream
34 informatics for improved long-read error correction and assembly. VirION2 assemblies combined with
35 short-read based data ('enhanced' viromes), provided significant improvements over VirION libraries in
36 the recovery of longer and more complete viral genomes, and our optimized error-correction strategy
37 using long- and short-read data achieved 99.97% accuracy. In the seawater virome, VirION2 assemblies
38 captured 5,161 viral populations (including all of the virus populations observed in the other assemblies),
39 30% of which were uniquely assembled through the inclusion of long-reads, and 22% of the top 10%
40 most abundant virus populations derived from assembly of long-reads. Viral populations unique to
41 VirION2 assemblies had significantly higher microdiversity means, which may explain why short-read
42 virome approaches failed to capture them. These findings suggest the VirION2 sample prep and workflow
43 can help researchers better investigate the virosphere, even from challenging low-biomass samples. Our

44   new protocols are available to the research community on protocols.io as a 'living document' to facilitate
45   dissemination of updates to keep pace with the rapid evolution of long-read sequencing technology.

46   **INTRODUCTION**

47   Microbes are recognized as a major driving force in the functioning and maintenance of most ecosystems
48   (Cavicchioli et al., 2019), however research in the past decade suggests viruses—mostly those that infect
49   bacteria ('phages')—are equally important. In the world's oceans, viruses (mostly dsDNA phages)
50   modulate microbial gene flow and are integral to global oceanic nutrient cycles (Brum et al., 2015; Roux
51   et al., 2016; Gregory et al., 2019). Because most viruses are uncultivated, advances have mainly arisen
52   through metagenomic approaches, which have rapidly improved with new sequencing technologies
53   (Brum & Sullivan, 2015). The importance of viruses in community composition and nutrient cycling is
54   also increasingly being recognized across diverse ecosystems, including soils (Emerson et al., 2018; Trubl
55   et al., 2018), the human microbiome (Shkoporov et al., 2019), glacial ice (Zhong et al., 2020), and
56   invertebrates (Shi et al., 2016; Wolf et al., 2020). For example, viruses have the potential to aid in soil
57   carbon flux by encoding plant polysaccharide-degrading enzymes (Emerson et al., 2018), or can be
58   involved in human gut dysbiosis that leads to various health issues (Mirzaei & Maurice, 2017).

59   Just as the scalability afforded by Illumina over 454 sequencing catapulted viromics from "gene ecology"
60   to "population-based or genome-resolved ecology" (Brum & Sullivan, 2015), long-read sequencing offers
61   promise for another transformative step forward (Warwick-Dugdale et al., 2019). Recent evidence
62   suggests that current short-read-sequencing metagenomic methods are not capturing the whole of the
63   virosphere. Within dsDNA viruses, the most extensively studied, short-read metagenomes capture
64   abundant viruses at the taxonomic level of species or genera, but not likely genotypes. This is because
65   complex virus communities contain mixtures of strains (i.e., virus population variants at the nucleotide
66   level), but assemblers cannot reconstruct individual strains —instead strain mixtures are collapsed into a
67   single 'consensus' genome or multiple genome fragments (Nurk et al., 2017), thus masking strain-specific
68   features that indicate the functional diversity within viral community members (Nelson et al., 2016).
69   Though the extent of populations missed is unknown, two separate studies – one using single-cell
70   genomics (Martinez-Hernandez et al., 2017) and the other long-read sequencing (Warwick-Dugdale et al.,
71   2019) – have demonstrated that strain-level diversity is under-represented in short-read datasets (with 3-
72   fold higher nucleotide diversity captured in long-read viral metagenomes compared to short-read viral
73   metagenomes; (Warwick-Dugdale et al., 2019)). Specifically, both studies demonstrate that high genome
74   microdiversity (i.e., the level of intra-population nucleotide variants for a virus 'species') could act as a
75   barrier to full genome assembly, while masking co-occurring population variants. Moreover, emerging
76   research suggests that hybrid assemblies (i.e., short- and long-reads co-assembly) could potentially
77   recover more genomic diversity in natural samples (Cook et al., 2020).

78   We previously introduced VirION (Warwick-Dugdale et al., 2019), a custom library sequencing and
79   informatics workflow that leveraged the strengths of both short- and long-read sequencing (using Oxford
80   Nanopore Technologies – 'ONT'). Briefly, this first entailed shearing of the sample DNA to ~8kb
81   fragments, followed by DNA amplification through 'Linker Amplified Shotgun Library' (LASL;
82   (Breitbart et al., 2002; Duhaime et al., 2012)) with a high-fidelity polymerase and 15 PCR cycles to
83   minimize rates of chimera formation. These steps produced sufficient DNA to meet the minimum DNA
84   input requirement (1 µg) for ONT sequencing. Long-read sequencing data was then assembled with an

85 overlap layout consensus strategy (via Canu; (Koren et al., 2017)), and contigs were error-corrected with
86 short-reads (with Pilon; (Walker et al., 2014)). Applying the full VirION workflow enabled a two-fold
87 increase in capturing 'complete' genomes and improve both the length and number of recovered niche-
88 defining hypervariable islands. However, the input DNA requirement (~100 ng extracted from 20 L of
89 seawater) for VirION is a barrier to generating long-read virome data from samples that yield little viral
90 DNA, either due to challenges in extraction and/or low-volume, high resolution sampling. Furthermore,
91 since the introduction of VirION, additional long-read assemblers and error-correction tools have been
92 introduced, which have not been evaluated for virus metagenomic datasets. Here, we introduce VirION2,
93 which includes significant wet-lab and analytical optimizations at key protocol steps to reduce input DNA
94 requirements and increase long-read lengths and accuracy. We applied this new workflow to a
95 concentrated natural seawater virioplankton community to assess VirION2's effectiveness at capturing
96 community features relative to VirION and other short-read approaches.

97 **MATERIALS AND METHODS**

98 *Phage mock community, DNA extraction and short-read sequencing*

99 Three *Pseudoalteromonas* phage isolates (PSA-HM1, PSA-HP1 and PSA-HS2; see **Supplementary**
100 **Table 1** for more details) were used as a mock community. Each phage was grown in culture and
101 genomic DNA extracted as previously described (Duhaime et al., 2017). To produce the mock
102 community, equimolar DNA concentrations from each phage extract were mixed. This mixture was used
103 as template for sequencing library preparation in all sequencing runs for optimizing VirION2 library
104 preparation. Separately, short-read sequencing libraries of the same mock community was prepared with
105 NEXTFLEX® Rapid DNA-Seq Kit 2.0 (Bio Scientific Corp, cat#NOVA-5188-01). Sequencing was
106 performed at the University of Exeter on a HiSeq 2500 instrument running the 2 x125 bp paired-end
107 configuration.

108 *Library preparations prior to Nanopore sequencing*

109 Three DNA library preparations were used: (1) Direct sequencing of DNA with no PCR amplification and
110 no shearing of viral DNA (referred to as 'unamplified'), followed by the manufacturer's sequencing
111 library protocol (using ligation sequencing kit SQK-LSK109; see below for sequencing library
112 preparation details). (2) The VirION pipeline, performed as previously described (Warwick-Dugdale et
113 al., 2019). Briefly, this entailed shearing of high-molecular weight DNA, followed by ligation of
114 amplification primers and PCR amplification of the sheared DNA prior to applying the manufacturer's
115 library protocol for sequencing. (3) VirION2, performed as follows: From the DNA extract, 10 μl were
116 set aside for short-read sequencing, and the remainder was sheared to 10 kbp or 15 kbp fragments with
117 Covaris g-TUBE according to the manufacturer's instruction (except for the 15 kbp treatment, in which
118 the samples were spun at 2,075 × *g* for 60 sec). Fragmented DNA was repaired and dA-tailed using the
119 NEBNext FFPE DNA Repair Mix and NEBNext Ultra II End Repair/dA-Tailing Module, according to
120 the manufacturer's instructions except incubation time, which was reduced from thirty to five minutes.
121 The repaired DNA was cleaned using Ampure Beads (Beckman Coulter) according to the manufacturer's
122 instructions (except for the final incubation time, in which 55°C was used instead of room temperature) at
123 a 1:1 ratio (v/v) and eluted in nuclease-free water. Ligation of barcoded adapters (Oxford Nanopore PCR
124 Barcoding Expansion 1-12 kit, cat# EXP-PBC001) used for linker-amplified shotgun library (LASL)

125   amplification was implemented as follows: A reaction mix composed of 50 µl NEB Blunt / TA Ligase
126   master mix, 20 µl of the Oxford Nanopore barcode, and 30 µl of cleaned DNA was incubated for ten
127   minutes at room temperature. The DNA was then immediately cleaned with Ampure beads with 1:0.4
128   sample to bead ratio to remove short fragments, and eluted in 15 µl nuclease-free water at 55°C. PCR
129   amplification on the cleaned libraries was performed as follows: A reaction mix using LA TaKaRa Hot
130   Start kit (Takara Bio USA) was prepared according to the manufacturer's instructions, using 5 µl of
131   cleaned DNA and 2 µl of the desired barcode (from the same Oxford Nanopore barcoding kit, as
132   previously). Cycling conditions were 94°C for 1 minute, 15 cycles of 94°C for 30 sec, 62°C for 30 sec,
133   68°C for 8-16 minutes, and final elongation for 8-16 min at 72°C. Elongation times varied according to
134   the desired amplicon size: 8 min for 10 kbp; 12 min for 15 kbp; and 16 min for 20 kbp. Amplicons were
135   subsequently cleaned with Ampure beads using a 1:0.5 sample to bead ratio, and eluted in 20 µl nuclease-
136   free water. The cleaned amplified libraries were then used as input for the 1D genomic DNA by ligation
137   kit (SQK-LSK109, Oxford Nanopore Technologies) according to the manufacturer's instructions, with
138   some modifications. In the 'DNA repair and end-prep' step, DNA CS (a standard DNA sequence used as
139   a positive/quality control by ONT) was excluded, and instead, 48 µl of input DNA was added. In addition,
140   incubation temperature was increased to 25°C from 20°C. In the 'Adapter ligation and clean-up' step, the
141   Long Fragment Buffer ('LFB') was used to enrich for longer fragments. In the bead resuspension step,
142   pellet resuspension was incubated for 10 minutes at 55°C (instead of room temperature) since better DNA
143   dissociation from the beads was observed to be more efficient at a higher temperature (data not shown).
144   The remainder of the protocol was unchanged.

145   *Mock community sequencing tests*

146   We tested four high-fidelity and/or long-range DNA polymerases in duplicate to evaluate improvements
147   of read length during the amplification step: 1) NEBNext (NEB M0541), 2) NEB Q5 (NEB M0491), 3)
148   NEB LongAmp (NEB M0287) and 4) TaKaRa LA Taq (TaKaRa RR042). For each library, 15
149   amplification cycles were used and input DNA (mock community phage genomes) was sheared at 10 kbp
150   (using Covaris g-TUBE). Influence of initial shearing length was also tested at 15 kbp with a DNA input
151   of 80 ng/µl. For TaKaRa LA Taq, we also tested a shearing length of 15 kbp using 1 ng of input DNA to
152   evaluate efficiency of VirION2 when minimal input DNA was available. For PCR cycling tests, 1 ng of
153   input DNA sheared at 15 kbp was used as template.

154   *Preparation of viral DNA from Western English Channel*

155   20 L of seawater from the Western English Channel ('WEC') was collected in rosette-mounted Niskin
156   bottles at a depth of 5m from the Western Channel Observatory (WCO;
157   http://www.westernchannelobservatory. org.uk/) coastal station 'L4' (50°15.00N; 4°13.00W) on 11th
158   February 2019. Seawater was placed in a coolbox at ambient temperature and upon return to shore, was
159   transported to the University of Exeter for processing within six hours of collection. The cellular fraction
160   was removed via sequential filtration through glass fiber (GF/D: pore size 2.7 µm) and polyethersulfone
161   (pore size 0.22 µm) filters in a 142 mm polycarbonate rig with a peristaltic pump. Viruses were
162   precipitated and concentrated from the filtrate by iron chloride flocculation and collected on 1.0 µm
163   polycarbonate filters (John et al., 2011). The viruses were immediately resuspended in ascorbate-EDTA
164   buffer (0.1 M EDTA, 0.2 M $MgCl_2$, 0.2 M ascorbic acid, pH 6.0) using 2 mL of buffer per 1 L of
165   seawater. The resuspended viral fraction was transferred equally to four Amicon Ultra 100 kDa

166   centrifugal filter units (Millipore UFC910024) which had been pre-treated with 1% bovine serum albumin
167   buffer to reduce capsid-filter binding (Deng et al., 2014) and flushed with SM buffer (0.1 M NaCl; 0.05
168   M Tris-HCl; 0.0008 M MgCl₂). The resuspended viral fraction was concentrated to 500-600 μl and
169   removed from the filter unit; the Amicon filters were then washed with 200 μl of SM buffer (Bonilla et
170   al., 2016) to ensure resuspension of all viral particles from capsid-filter adhesion. The viral fraction was
171   purified with DNase 1 (100 U/mL; 2 hours at room temperature) to remove non-encapsulated DNA.
172   DNase 1 activity was terminated by the addition of 0.1 M EGTA and 0.1 M EDTA (Hurwitz et al., 2013).
173   Viral DNA was extracted from the concentrated and purified viral fraction using Wizard DNA Clean-Up
174   System (Promega A7280) to remove PCR inhibitors (John et al., 2011). The viral DNA was cleaned and
175   concentrated by a 1.5✕ Ampure bead cleanup for downstream application.

176   *Long and short read processing of Western English Channel viral DNA*

177   Viral DNA from the WEC was used to prepare three separate long-read sequencing libraries and one
178   short-read only library for comparison and error correction. (1) *VirION libraries* were prepared as
179   described previously (Warwick-Dugdale et al., 2019), using 100 ng of input DNA; (2) *Unamplified*
180   sequencing of viral DNA was performed using ~3 μg of unsheared input DNA in a LSK-SQK109 library
181   preparation (ONT) for genome sequencing according to the manufacturer's instructions, and sequenced
182   on a MinION R9.4 revD flowcell for 48 hours. (3) A *VirION2* library was prepared by first treating the
183   DNA with the Zymo DNA Clean & Concentrator Kit (cat# D4013) according to the manufacturer's
184   instructions. The clean concentrate (90 μl, final concentration of 14.6 ng) was then sheared into 15 kbp
185   fragments, used as template for 15 PCR cycles, and processed for MinION sequencing on a single
186   flowcell according to the manufacturer's instructions with the modifications stated in the previous section
187   ("*Library preparations prior to Nanopore sequencing*"); (4) *Short Read sequencing* was performed using
188   51 ng of input DNA in a 1S Plus (Swift Biosciences) library preparation and sequenced to a depth of
189   ~25M 2x125bp paired end sequencing on a HiSeq 2500 at the University of Exeter.

190   *Long-read processing and quality control*

191   Raw fast5 files containing ONT long-reads were basecalled with Guppy v2.3.1 (provided by Oxford
192   Nanopore), using the flip-flop model. Reads were allocated into barcode bins by porechop (Wick, 2017),
193   using --require_two_barcodes, --discard_unassigned and --discard_middle parameters in order to limit
194   library cross-talk in multiplexed samples, and to remove concatenated reads where two strands pass
195   through the same pore in quick succession. NanoFilt 2.2.0 (De Coster et al., 2018) was used to trim the
196   first 50 bases of reads (to ensure any residual barcode sequence removal) and remove reads < 1kbp or
197   with a phred quality score < 9. Chimeric PCR products were identified for quantification and subsequent
198   removal from each demultiplexed run using yacrd (Marijon, Chikhi & Varré, 2020) with default
199   parameters. Remaining reads were kept for further analysis. To compare read lengths between mock
200   community datasets, we randomly subsampled the reads (with replacement) using the 'sample' package in
201   R version 3.5.0. Differences in bootstrapped medians (n=50,000, 1000 replicates) between the barcoded
202   libraries associated with each treatment (DNA polymerase type, DNA shearing size and PCR cycling
203   number) were plotted as boxplots with the R package 'ggpubr'.

204   *Assessing assembly and error profiles*

205 For testing assembly performance of the WEC sample, reads that passed quality control (described above)
206 were first sub-sampled using bbtools reformat (Bushnell, 2015) in order to mitigate library size bias in
207 assembly comparisons. Sub-sampling was based on the number of reads in the smallest library (500,000
208 reads in the unamplified dataset). We compared median read length between the full datasets and the sub-
209 sampled datasets to confirm subsampling had not biased read length distributions (**Supplementary**
210 **Figure 1**), and observed negligible shifts between full and sub-sampled datasets in both VirION (median
211 read length: 3,020-3,023 bp 95% CI and 3,018-3,027 bp 95% CI) and VirION2 datasets (3,906-3,916 bp
212 95% CI and 3,900-3,917 bp 95% CI). Subsampled reads were used in three assembly approaches to
213 determine optimal assembly strategy: (1) *Overlap-layout consensus assembly* ('OLC'): All-vs-all
214 alignments were generated using minimap2 v2.17-r941 (Li, 2018), and used to build an assembly with
215 Miniasm v0.3-r179 (Li, 2016). Minipolish v0.1.2 (Wick & Holt, 2020) was used to iteratively apply
216 Racon-based contig polishing (Vaser et al., 2017) on Miniasm assemblies. Polished OLC assemblies in
217 GFA format were converted to FASTA using the following awk command: " '$1 ~/S/ {print
218 ">"$2"\n"$3}' assembly.gfa > assembly.fasta". Fasta-formatted Racon assemblies were further error-
219 corrected with Medaka v0.11.5 (Oxford Nanopore Technologies Ltd., 2018), with an appropriate model
220 select to correspond to the sequencing chemistry and basecaller model used. Lastly, Pilon v1.23 (Walker
221 et al., 2014) was used to correct errors using short-read mapping information to the polished long-read
222 assemblies. Briefly, short reads were mapped using BWA-MEM v0.7.17-r1198 (Li, 2013) to the Medaka-
223 corrected assemblies, and the resulting bam files were used in a single iteration of Pilon correction (using
224 --fix all). (2) *Flye assembly*: Reads were first assembled into unitigs using the --nano-raw and --meta
225 parameters in Flye v2.5 (Kolmogorov et al., 2019) , with an estimated assembly size of 15M. Medaka and
226 Pilon were then used to correct Flye assemblies as previously described. (3) *Hybrid SPAdes assembly*:
227 Hybrid assemblies were generated with hybridSPAdes v3.12.0 (Antipov et al., 2016), using the --meta
228 and --nanopore parameters. For comparison and evaluation of error in long-read assemblies, short-read
229 only assemblies were assembled with metaSPAdes v3.12.0 (Nurk et al., 2017) with default parameters.
230 The error-correction performance (number of mismatches and indels) of each long-read assembly/error-
231 correction strategy (and intermediary stages) was assessed with Quast v4.5 (Gurevich et al., 2013) using
232 the -*meta* option with default parameters against the matching short-read only assembly of the same WEC
233 sample. Genome completeness was computed using CheckV v0.3.0 (Nayfach et al., 2020). Per contig
234 median microdiversity (π) was computed from single nucleotide polymorphism (SNP) frequencies as
235 described previously (Warwick-Dugdale et al., 2019). Briefly, short-reads were first mapped to viral
236 contigs using BWA-MEM with mapping thresholds (70% read coverage at 95% identity) previously
237 established (Gregory et al., 2019) to define viral populations. Of these mapped genomes, only those that
238 had >70% of their length covered by reads, and had >10x coverage were further selected. For each viral
239 population, SNPs were identified using mpileup within samtools (Li et al., 2009) and BCFtools (Danecek
240 & McCarthy, 2017), and low quality variant call (PHRED<20) were removed. SNPs were identified as
241 'true' if an alternate variant nucleotide was present on at least four mapped reads, and comprised >1% of
242 the base pair coverage for each variant nucleotide at that position..

243 *Generating 'enhanced' viromes with short- and long-read assemblies*

244 To produce 'enhanced' virome datasets, each assembly type (i.e., metaSPAdes, hybridSPAdes, long-read
245 'OLC') were independently searched for virus contigs (contigs ≥2.5kbp) using VirSorter v1.0.5 (Roux et
246 al., 2015) in --virome search mode. In each assembly type, only viral contigs ≥5kbp belonging to
247 VirSorter categories 1, 2, 4, and 5 were kept. Filtered virus contigs from each assembly were pooled and

248   subsequently dereplicated into viral populations with ClusterGenomes (Roux, 2015), using 70% coverage
249   and 95% nucleotide identity clustering thresholds (*sensu* (Brum et al., 2015)).

250   **RESULTS AND DISCUSSION**

251   *Experimental overview*

252   Two sets of experiments were performed (**Figure 1**) to evaluate and optimize our long-read laboratory
253   protocol in terms of increased read length, reduced DNA input requirements, and minimized error in
254   assembled genomes. *Experiment 1* sought to maximize read length and minimize chimera formation from
255   sheared, amplified DNA extracted from a mock community of cultured phages. This community
256   contained a representative member of the main families of tailed phages (*Myoviridae*, *Siphoviridae*, and
257   *Podoviridae*) and ranged in genome sizes (38.2 kbp - 129.4 kbp) and GC content (35.7% - 44.7%) (see
258   Methods, **Figure 1A**, and **Supplementary Table 1**). We evaluated four DNA polymerases (*Experiment
259   1A*) using input material sheared to 10kbp; optimized DNA shearing size to generate long-read viromes
260   from a natural seawater sample (*Experiment 1B*), and evaluated the influence of the number of PCR
261   cycles of chimera formation and read length (*Experiment 1C*). In *Experiment 2* we optimized informatic
262   approaches to maximize viral genome recovery and reduce error using the long-read virome datasets from
263   Experiment 1B in combination with short-reads. We evaluated two assembly/error-correction strategies
264   (*Experiment 2A*) and benchmarked these against assemblies from short-read approaches, which were
265   assumed to have negligible sequencing error (*Experiment 2B*).

266   *Experiment 1A: Amplification with TaKaRA LA yields highest median read lengths*

267   Among the four polymerases tested, amplification with TaKaRa LA Taq yielded the highest median read
268   length (6,965 bp; 6,957 – 6,973 bp 95% CI). This was a significant increase (p < 0.0001) of 2,866 bp
269   compared to the performance of NEBNext used in the VirION workflow (Warwick-Dugdale et al., 2019),
270   but retained a low proportion of chimeric reads (0.08-0.19%). Therefore, we proceeded with TaKaRa LA
271   Taq for all subsequent tests.

272   *Experiment 1B: Longest read lengths are achieved at low DNA concentrations sheared to 15kbp*

273   We next sought to evaluate if increased fragment length of 15 kbp provided a concomitant increase in
274   final read length compared to 10 kbp, and whether this was also compatible with low-input DNA
275   concentrations typical of some environmental viromes. Previous work using VirION had shown a
276   discrepancy between expected fragment length (~8 kbp) and final median read lengths (~4.1 kbp),
277   suggesting that longer fragments do not necessarily yield significantly longer reads. Final read length may
278   be constrained by the amplicon size produced by the DNA polymerase or preferential amplification of
279   smaller fragments (Shagin et al., 1999; Warwick-Dugdale et al., 2019). Here, shearing 80 ng of input
280   DNA to 15 kbp reduced final median read lengths (6,240 bp; 6,238 – 6,242 bp 95%CI) in comparison to
281   the 10 kbp shearing treatment (median: 6,965 bp; see Experiment 1A) (**Figure 2B**). However, the longest
282   median read lengths were observed when 1 ng of input DNA was sheared to 15 kbp (median: 7,223 bp;
283   6,899 – 7,540 bp 95% CI), an increase of 216 bp and 983 bp over shearing 80 ng DNA to 10 kbp and 15
284   kbp, respectively. There was significant variance in read length between replicates (**Figure 2B**),
285   suggesting that: (1) final read lengths may be strongly influenced by variance in downstream library

286 preparation flowcell properties; (2) Interactions between input DNA and centrifugal force selected for
287 shearing may be complex. Both DNA inputs <100 ng and a shear size of 15kbp are outside the official
288 parameters of Covaris g-tube specifications. Therefore, additional optimizations may be required to
289 reduce variance between samples.

290 To assess our revised protocol on real world samples, three long-read libraries (from 14.6 ng of input
291 DNA) from a virus-enriched marine sample were generated and sequenced (**Figure 2C**). These libraries
292 were (1) a 'no amplification' library ('unamplified'), (2) our previous VirION protocol, and (3) this study
293 ('VirION2'). The 'unamplified' library had the highest median length (5,601 bp; 5,560 – 5,645 bp 95%
294 CI). Between the amplified datasets, the read size distribution from the VirION2 library was significantly
295 higher than VirION. However, in both libraries read sizes were generally lower than those observed in the
296 mock community (at identical shearing size), a phenomenon also observed in our previous study
297 (Warwick-Dugdale et al., 2019). This is likely due to shearing and/or degradation during the $FeCl_3$
298 precipitation, giving smaller fragments for amplification in the natural sample. Further optimization into
299 maximizing DNA integrity from viral metagenomes could be beneficial to improving long-reads.

300 *Experiment 1C: VirION2 libraries can be prepared with very low input DNA and increased PCR cycles*
301 *without increasing chimeric reads*

302 Samples with low DNA concentrations require increased numbers of PCR cycles to meet the input
303 requirements (1 μg) of Nanopore sequencing. To maximize VirION2's applicability to various sample
304 types, including those with very low input DNA concentrations (~1 ng), we next tested how amplification
305 cycling (15, 18, 20, and 22 cycles with TaKaRa LA Taq) impacted read lengths and chimeras (**Figure
306 2D**). Starting with the mock community DNA at 1ng concentration, the 20-cycle treatment generated
307 significantly longer reads compared to all other treatments (Wilcoxon rank test, $p < 2.22 \times 10^{-16}$; median =
308 8,125 bp). Across all treatments and replicates in this experiment, the number of chimeric reads remained
309 near constant (range: 0.01 – 0.04%) and therefore, at least within the range tested here, is not an obstacle
310 to increasing cycle number to generate sufficient DNA for sequencing.

311 *Experiment 1 conclusions*

312 Together, the optimization of experiment 1 improve the VirION method in two critical ways. First, the
313 median read length was significantly increased (p < 0.0001) by 76% over the original VirION approach.
314 Second, we were able to generate long-read data from 1ng libraries with a negligible number of chimeric
315 reads, which removed the one *micro*gram requirement for standard MinION libraries, and reduced
316 VirION's DNA input requirement 100-fold (Warwick-Dugdale et al., 2019). These laboratory protocol
317 improvements should permit a broader diversity of samples to be investigated with long-read sequencing
318 technology. Since Nanopore sequencing is constantly evolving, we have posted the VirION2 method at
319 protocols.io to ensure continued protocol development as the research community identifies new
320 opportunities.

321 *Experiment 2 – Informatics benchmarking*

322 In a second experiment (**Figure 1B**), we assessed two informatic approaches for combining long- and
323 short-read data to maximize recovery and accuracy of viral genomes from natural viral communities. A

324   single, virus-enriched seawater sample ('WEC'; see Methods and Experiment 1B) was used, along with
325   the corresponding Illumina short-reads. Adding short reads served two purposes: 1) as a gold standard for
326   long-read sequencing accuracy estimates and 2) for use in hybrid assemblies and error correction of long-
327   read assemblies.

328   *Experiment 2A: Performance of long-read assemblers and error-correction tools.*

329   We first tested an 'overlap-layout consensus' ('OLC strategy') assembly approach (performed by
330   Miniasm), followed by the successive application of three error-correction tools: Racon, Medaka, and
331   Pilon (the latter uses short-read correction). Of note, assembly of long-read data from this study with
332   Canu (Koren et al., 2017), used in VirION (Warwick-Dugdale et al., 2019) failed to finish running within
333   48h wall time on an high-performance computing node with 48 CPU cores and 1Tb RAM (data not
334   shown), likely due to the high volume of data generated in contemporary MinION runs (Amarasinghe et
335   al., 2020). In the 'OLC strategy', Racon was used to correct raw contigs by generating genomic consensus
336   through multiple iterations of long-read self-mappings, however there is no set standard for how many
337   rounds of Racon to use. Therefore, we first assessed the impact of Racon iterations on error correction by
338   testing 1, 2, 3, 5 and 10 rounds of Racon polishing. Two rounds of polishing were sufficient to produce a
339   sharp decrease (2 – 3-fold drop) into the number of mismatches (per 100 kbp of sequence), while a single
340   round resulted in a 5 - 8.5-fold reduction in insertion/deletion ('indels') events (**Supplementary Figure**
341   **2A**). Further rounds of polishing made no significant difference to the number of mismatches and indels
342   removed. Concomitantly, a single round of polishing significantly improved median predicted protein
343   lengths from 71 a.a to 107 a.a but additional rounds resulted in no significant improvements
344   (**Supplementary Figure 2B**). Next, we estimated the effectiveness of error correction between three
345   types of long-read libraries prepared from the same WEC sample (i.e., 'unamplified', 'VirION and
346   'VirION2'), compared to short-read only assemblies. Applying Medaka and Pilon provided a moderate
347   reduction (range: 2-13%) of mismatches and a major reduction in indels (range: 70-73%; **Supplementary**
348   **Figure 3A**) across all three library types. Short-read error correction of long-read assemblies with Pilon
349   had the greatest impact in partially restoring the expected median protein size of 142 a.a predicted from
350   short-read only data (dotted line, **Supplementary Figure 3B**). This was most effective in the VirION2
351   library, reaching 131 a.a median size (130 – 132 a.a 95% CI), compared to 113 a.a (112 – 115 a.a 95%
352   CI) in VirION and 112 a.a (111 – 113 a.a 95% CI) in the 'unamplified' dataset. Therefore OLC assembly
353   of VirION2 reads, coupled with two rounds of Racon polishing, followed by one round of Medaka and
354   additional short-read polishing with Pilon is our recommended approach, and yielded the most improved
355   assembly correction between the three long-read libraries, as well as reaching median protein size closest
356   to the short-read assembly.

357   We next tested the effectiveness of assembly by repeat graphs using Flye (see Methods) to process our
358   virome. Recent reports (De Maio et al., 2019; Moss, Maghini & Bhatt, 2020) suggest good performance
359   in terms of assembly accuracy and scale in prokaryotic datasets, however its performance with viromes
360   remains untested. Assembly of WEC long-read datasets with Flye was followed by the same error-
361   correction tools as described for 'OLC strategy', except for Racon (Racon is designed for Miniasm only,
362   and Flye has an internal error-correction module). Despite repeated attempts and large computing
363   allowances (48 CPU cores/1Tb RAM), we could not generate a Flye assembly from the 'unamplified'
364   library. This could be due either to sequence complexity in the sample, which can complicate estimating
365   the required genome size parameter of the program, and/or insufficient resources to run (Wick & Holt,

366   2020). However, both VirION and VirION2 datasets were successfully assembled by Flye without large
367   resources, therefore we focused our analysis on these. Similar to the OLC strategy, indel reduction was
368   more successful (range: 74-75% removed) compared to mismatches. Surprisingly, mismatches in Flye-
369   assembled contigs marginally increased with polishing by Medaka – a feature that was not entirely
370   corrected with subsequent application of short-read polishing with Pilon, resulting in an overall 1.25%
371   removal in VirION2 only, but an increase of 4.53% in VirION) (**Supplementary Figure 3C** and
372   **Supplementary Table 2**). Corrected Flye assemblies had shorter median protein sizes compared to the
373   OLC strategy results in both VirION and VirION2 datasets (by 5 and 21 amino acids, respectively,
374   **Supplementary Figure 3D**), perhaps due to greater indel reduction in the OLC strategy. Therefore, in
375   terms of correction performance, the OLC strategy is preferred over the Flye strategy, at least for
376   assembly of long-read viral metagenomics with current software versions.

377   *Experiment 2B: VirION2 coupled with OLC assembly provides greatest gains in assembly and predicted*
378   *gene lengths*

379   Using the results generated in experiment 2A, we compared all the corrected long-read assembly types to
380   hybrid assemblies (i.e., short read assemblies scaffolded by long-reads) and short-read only assemblies
381   (**Figure 3**). Across all assembly types (hybrid, OLC and Flye) and libraries (unamplified, VirION and
382   VirION2), sequence accuracy of polished long-read assemblies ranged from 99.61% to 99.98%
383   (**Supplementary Table 2**). The VirION2 OLC strategy offered an increased accuracy of ~0.05% over
384   'unamplified' and VirION-amplified reads. Although seemingly small, this increased accuracy translated
385   into a much-improved median protein size of 131 a.a (130 – 132 a.a; 95% CI), compared to 112 a.a (111 –
386   113 a.a; 95% CI) in the unamplified library and 113 a.a (112 – 115 a.a; 95% CI) for the VirION library.
387   Regardless of library method, Flye assemblies consistently yielded the lowest median protein sizes. In
388   terms of contigs lengths, N50 metrics were markedly improved (~4-fold) in all assembly strategies in
389   comparison to short-read only assemblies (**Supplementary Table 3**). Between VirION and VirION2
390   specifically, the N50 from VirION2 contigs was superior in both OLC (31,496 bp versus 16,840 bp) and
391   Flye strategies (31,130 bp versus 21,231 bp). The same was true for maximum contig sizes reached:
392   194,588 bp versus 115,168 bp (OLC), and 570,045 bp versus 140,205 bp (Flye). Overall, between library
393   preparation (unamplified, VirION and VirION2) and long-read assembly/correction strategies (OLC and
394   Flye), combining VirION2 with the OLC strategy performed best (99.71% accuracy, **Supplementary**
395   **Table 2**) and is the preferred option for accurate long-read virome datasets, with the critical advantage of
396   lowering the input DNA requirement by 100-fold.

397   *VirION2-enhanced viromes recovers longer and more microdiverse viral populations*

398   Next, we evaluated how well short- and long-read assemblies recovered viral populations from a natural
399   viral community, in terms of genome size, genome completeness and microdiversity. Specifically, we
400   compared a short-read-only virome against 'enhanced' viromes that maximize the benefits of both short-
401   and long-read technologies (derived from combining multiple assemblies from both read types; see
402   Methods and **Figure 4A**). Genome size, genome completeness, and microdiversity (i.e., intra-population
403   genomic variation) were all significantly improved in VirION2-enhanced viromes, compared to short-
404   read and VirION-enhanced viromes (**Figure 4 B –D**, and **Supplementary Table 4**). Notably, recovery of
405   >100 kbp genomes in the VirION2-enhanced virome was increased by 13- and 2-fold (**Figure 4B**),
406   relative to short-reads only and VirION-enhanced, respectively. Accurate and contiguous assembly of

407 large phage genomes could increase the detection of 'jumbo' phages throughout ecosystems, as there is
408 currently limited data on the distribution and genome diversity for large phages (Yuan & Gao, 2017; Al-
409 Shayeb et al., 2020). Microdiversity (measured as π, **Figure 4D**) increased by 2.4 - 2.5-fold in both
410 VirION- and VirION2- enhanced viromes, respectively, compared to the short-read assemblies. A similar
411 increase was reported in the original VirION method from WEC seawater samples (3-fold; (Warwick-
412 Dugdale et al., 2019)). However, there was no significant difference between VirION and VirION2
413 datasets (Mann Whitney U-test, *p-value* = 0.536). Although it is tempting to speculate that residual
414 sequencing error in long-read datasets could artificially increase SNP frequency and therefore higher
415 microdiversity (π) values, we found that its impact is negligible due to the error-correction pipeline
416 applied, along with the error distribution patterns (mostly indels) of Nanopore reads (for more details, see
417 **Supplementary Analysis 1**). Alternatively, the higher microdiversity in 'enhanced' datasets compared to
418 short-read data could be due to novel viral populations uniquely captured with long-reads (discussed
419 below). In addition, it is worth noting that during the multiple rounds of polishing of long-read
420 assemblies, genuine biological diversity may be removed from consensus sequences, leading to an
421 underestimation of microdiversity.

422 *Long-read viromics improves capture of abundant viral populations compared to short-read only*
423 *methods*

424 In addition to better genomic metrics, we assessed how the 'VirION2-enhanced' assembly from the WEC
425 seawater sample could alter our biological interpretations of complex viral communities. We first
426 evaluated if more viruses could be detected by adding long-reads, and by counting the number of shared
427 and unique viruses in each constituent assembly that make the VirION2-enhanced virome (i.e., long-read
428 only, short-read only, and hybrid). Among the total number of viral populations (n=5,161), 68% were
429 represented in all assembly types. Within the pool of viral populations observed only within a single
430 assembly method, populations from the hybrid assembly were the majority (23%), followed by the OLC
431 assembly (7%), and short read assembly (2%). We surmise that adding long reads permitted a more
432 complete view of a virome, and viral genomes that would otherwise have been missed by short-read
433 assembly alone.

434 Next, we looked at patterns in abundance of viral populations as a function of assembly method within the
435 VirION2-enhanced virome. (**Figure 5A**). Among all viruses, the majority of these unique viruses were
436 low abundance populations, and were short fragments from the hybrid assembly. However, within the top
437 10% most abundant viruses (n=516), 164 were unique to a particular assembly method and 70% of these
438 were derived from the long-read only assembly. Taken together, the addition of long-reads uniquely
439 enabled the recovery of 22% of the most abundant viruses—that would have been missed in short-read
440 only assemblies. We queried whether microdiversity could be a contributing factor as to why these unique
441 viruses would be missed. Indeed, when we looked at the fraction of 'unique' viral populations within each
442 assembly type in the 'enhanced' virome, unique viral populations had significantly higher microdiversity
443 compared to the shared pool of viruses in each assembly, most predominantly observed in the 'OLC'
444 assembly (**Figure 5B**). Thus, our data suggest that high intrapopulation nucleotide variation in a subset of
445 virus genomes can lead to these genomes being fragmented (Roux et al., 2017) and overlooked in a short-
446 read only viral metagenome experiment.

447 *Current limitations and future directions*

448 Despite the advances of VirION2, several limitations remain. First, an assembly step (due to the required
449 shearing of input DNA), as well as corresponding short-reads, remain a necessity to ensure the highest
450 level of error-correction, and to enable microdiversity estimates. Ongoing developments that increase
451 recovery of full-length genomes as single long-reads and optionally use short-reads for correction
452 (Beaulaurier et al., 2020) will certainly help. However, this workflow cannot be used for low-input
453 samples, as it requires micrograms of high molecular weight DNA to yield full genomes as single reads.
454 Moreover, even if sufficient DNA is obtained, our tests using 'unamplified' datasets indicated that
455 compared to the VirION2 approach, viral populations from the unamplified library had shorter median
456 proteins sizes (112 a.a versus 131 a.a, respectively). Second, the revised long-read library protocol was
457 designed and evaluated for the recovery of dsDNA virus genomes. Future refinements could include
458 modifying the current protocol for the inclusion of ssDNA and RNA genomes (through direct RNA or
459 cDNA libraries). Third, due to the amplification step within the VirION2 laboratory protocol, we cannot
460 leverage base modification capabilities from Nanopore sequencing, as these can only be detected if intact,
461 non-amplified high molecular weight DNA is used as template for sequencing. Lastly, we have not
462 evaluated VirION2 on the latest R10 flowcells, which promises further sequencing accuracy, especially
463 within homopolymeric regions. Adapting VirION2 to R10 chemistry (R9.4 flowcells were used in this
464 study) to generate long-read viromes, as well as sequencing depth available from the PromethION (the
465 higher throughput version of the MinION sequencing platform) remain to be tested, but we do not
466 anticipate major protocol (either wet-lab or informatics) modifications will be required. Lastly, additional
467 laboratory optimizations parameters could be tested, including whether further increases in DNA shearing
468 sizes (e.g., 20kbp or higher) may further improve long-read sizes, albeit with likely diminishing returns.

469 **Conclusions**

470 Though short-read sequencing has become the gold standard in viral metagenomics, it is increasingly
471 clear that this approach does not capture the full extent of virus macrodiversity (i.e., species richness) and
472 microdiversity (i.e., intra-population genomic diversity). The latter in particular, can cause genome
473 fragmentation and mask genes under active selection pressures (e.g. genomic islands) and important
474 indicators of virus-host dynamics. Long-read sequencing, combined with short-reads, can further increase
475 estimates of viral diversity and capture ecologically important taxa. Owing to the substantial decrease in
476 DNA requirements for VirION2, long-read viral metagenomics may now be applied to a broader range of
477 samples, thus constituting an invaluable addition to our current viral ecogenomics toolkit for the better
478 exploration of viromes and their impacts in nature.

479 **Acknowledgements**

480 The authors thank the crew of the Plymouth Marine Laboratory vessel 'Quest' for collection of seawater
481 samples. Research was supported by high performance computing resources provided by the University
482 of Exeter and the Ohio Supercomputer Center.

483 ***Figure and Table legends***

484 **Figure 1.** Overview of wet lab optimization experiments and informatic benchmarking. **A)** Laboratory
485 optimization ('Experiment 1') in which a mock community of three phages was used to conduct three
486 experiments aimed at producing longer reads from less input DNA. **B)** Informatics benchmarking

487   ('Experiment 2') in which a seawater virome was sequenced with short-reads (Illumina) and long-read
488   sequencing. Three distinct long-read libraries were generated, error-corrected and assembled, and were
489   compared to short-read assemblies to assess accuracy and assembly performance.

490   **Figure 2.** Laboratory optimization yield longer reads from less DNA. **A)** Boxplots showing the median
491   and quartiles of the read length distribution between four DNA polymerases. **B)** Boxplots showing the
492   median and quartiles of the read length distribution between DNA shearing size treatments and one low-
493   input DNA (1 ng) variant of the 15 kbp shearing treatment. **C)** Boxplots showing the median and quartiles
494   of the read length distribution of three long-read library types, either unamplified or amplified (VirION
495   and VirION2). **D)** Boxplots showing the median and quartiles of the read length distribution between four
496   thermocycling treatments (here, number of cycles). Asterisks represent a significant difference
497   ($p<0.0001$) between pairs of replicate treatments where applicable.

498   **Figure 3.** Error-correction profiles between library methods and assembly strategies using the WEC
499   sample. **A)** On the x-axis, mismatches and insertion/deletion ('indels') events according to assembly
500   strategy (full OLC, full Flye, and Hybrid) are grouped separately, and divided into three facets, one for
501   each long-read library method (unamplified, VirION, VirION2). The number of errors (y-axis) is scaled
502   to the binary logarithm (log2) for scale fitting purposes. **B)** Boxplot depicting the protein size distribution
503   (in amino acids, denoted as 'a.a'; y-axis) derived from each library method (x-axis), each of which is sub-
504   grouped per assembly strategy. In both A and B panels, there were no results from Flye assemblies for the
505   raw datasets, as these could not be produced.

506   **Figure 4.** Comparison of virus genome properties between short-read and 'long-read-enhanced' viromes.
507   **A)** Workflow to produce 'enhanced viromes', in which Spades, hybrid and long-read (OLC) viruses are
508   combined to maximize the recovery of virus signals. **B)** Cumulative Distribution Function (CDF) plot
509   depicting the frequency (y-axis) of virus genomes according to genome length (measured in kilo basepairs
510   (kbp), x-axis) between three assembly strategies. **C)** Cumulative Distribution Function (CDF) plot
511   depicting the frequency (y-axis) of virus genomes according to genome 'completeness' (measured in %,
512   x-axis) between three assembly strategies. **D)** Cumulative Distribution Function (CDF) plot depicting the
513   frequency (y-axis) of virus genomes according to genome microdiversity per genome (measured as π, x-
514   axis) between three assembly strategies.

515   **Figure 5.** Additional insights are gained through a VirION2-enhanced assembly strategy. **A)** Rank
516   abundance curve depicting the seawater virus community, colored according to whether a virus
517   population (individual bars) was detected uniquely (turquoise) or in multiple (pastel red) assembly types.
518   The top 10% most abundant viral populations are highlighted between dashed lines, where they are
519   divided per assembly origin in the pie chart. **B)** Boxplots depicting the level of microdiversity between
520   shared and unique viral populations within each constituent assembly present in the 'enhanced' dataset.

521   **Supplementary Figure 1**. Visualizing sub-sampling impacts on long-read size datasets. Frequency
522   histograms depicting the read size distribution between the three long-read libraries: 'raw' (i.e.,
523   unamplified input DNA), VirION, and VirION2. Each library type is represented in its own facet, with
524   both full (pastel red) and sub-sampled (turquoise) size distributions overlapping. Dotted lines and their
525   associated colors denote the median value of each distribution.

526 **Supplementary Figure 2**. Impact of error-correction with the Racon tool using VirION2 data from a
527 seawater sample ('WEC'). **A)** Bar chart depicting the number of mismatches and insertion/deletions
528 (indels), grouped by number of Racon rounds (on the x-axis) across increasing iterations of Racon
529 polishing of long-read assemblies. From light to dark blue, increasing color saturation corresponds to the
530 increasing number of Racon polishing rounds (also applicable to panel B). **B)** Boxplot depicting the
531 impact of consecutive rounds of Racon (x-axis) on predicted protein sizes (y-axis, measured in amino
532 acids denoted as 'aa'). The horizontal dotted line represents the median protein size of the corresponding
533 short-read assembly (142 amino acids).

534 **Supplementary Figure 3**. Error-correction levels at each correction stage within OLC and Flye strategies
535 compared between long-read library methods. **A)** Bar chart depicting the number of mismatches and
536 insertion/deletions (indels) from the OLC strategy, grouped according to increasing correction (on the x-
537 axis), separated by library method (raw, VirION, VirION2). **B)** Boxplots depicting protein size
538 distribution at each level of correction in the OLC strategy. The horizontal dotted line represents the
539 median protein size of the corresponding short-read assembly (142 amino acids). **C)** Bar chart depicting
540 the number of mismatches and insertion/deletions (indels) from the 'Flye' strategy, grouped according to
541 increasing correction (on the x-axis), separated by library method (VirION and VirION2). **D)** Boxplots
542 depicting protein size distribution at each level of correction in the 'Flye' strategy. The horizontal dotted
543 line represents the median protein size of the corresponding short-read assembly (142 amino acids).

544 **Supplementary Figure 4**. Per-genome microdiversity distributions within the VirION2-enhanced
545 viromes constituent assemblies.

546 **Supplementary Table 1.** Summary characteristics of the virus isolates used in the mock community for
547 Experiment 1.

548 **Supplementary Table 2.** Error-correction performance benchmarks between assembly methods and
549 sequencing libraries.

550 **Supplementary Table 3.** Contigs and protein statistics between assembly strategies above 2.5 kbp.

551 **Supplementary Table 4.** Genome-based metrics between Nextera-only and virION-enhanced datasets
552 (>5 kbp genomes).

553 **Supplementary Table 5.** Tracking SNP and microdiversity detection in mock community long-read
554 assemblies at different stages of error correction.

555 **References**

556 Al-Shayeb B, Sachdeva R, Chen L-X, Ward F, Munk P, Devoto A, Castelle CJ, Olm MR, Bouma-
557 Gregson K, Amano Y, He C, Méheust R, Brooks B, Thomas A, Lavy A, Matheus-Carnevali P, Sun
558 C, Goltsman DSA, Borton MA, Sharrar A, Jaffe AL, Nelson TC, Kantor R, Keren R, Lane KR,
559 Farag IF, Lei S, Finstad K, Amundson R, Anantharaman K, Zhou J, Probst AJ, Power ME, Tringe
560 SG, Li W-J, Wrighton K, Harrison S, Morowitz M, Relman DA, Doudna JA, Lehours A-C, Warren
561 L, Cate JHD, Santini JM, Banfield JF. 2020. Clades of huge phages from across Earth's ecosystems.
562 *Nature* 578:425–431. DOI: 10.1038/s41586-020-2007-4.

563 Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in
564   long-read sequencing data analysis. *Genome Biology*. DOI: 10.1186/s13059-020-1935-5.

565 Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. HybridSPAdes: An algorithm for hybrid
566   assembly of short and long reads. *Bioinformatics* 32:1009–1015. DOI:
567   10.1093/bioinformatics/btv688.

568 Beaulaurier J, Luo E, Eppley JM, Uyl P Den, Dai X, Burger A, Turner DJ, Pendelton M, Juul S,
569   Harrington E, DeLong EF. 2020. Assembly-free single-molecule sequencing recovers complete
570   virus genomes from natural microbial communities. *Genome Research* 30:437–446. DOI:
571   10.1101/gr.251686.119.

572 Bonilla N, Rojas MI, Netto Flores Cruz G, Hung S-H, Rohwer F, Barr JJ. 2016. Phage on tap–a quick and
573   efficient protocol for the preparation of bacteriophage laboratory stocks. *PeerJ* 4:e2261. DOI:
574   10.7717/peerj.2261.

575 Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002.
576   Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of
577   Sciences of the United States of America* 99:14250–14255. DOI: 10.1073/pnas.202488399.

578 Brum JR, Ignacio-espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cesar Ignacio-
579   Espinoza J, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, De Vargas C, Gasol
580   JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos
581   BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Bork P, Bowler
582   C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB, Boss E, Follows M, Grimsley N, Jaillon O,
583   Karp-Boss L, Krzic U, Raes J, Reynaud EG, Sardet C, Sieracki M, Stemmann L, Velayoudon D,
584   Weissenbach J. 2015. Patterns and ecological drivers of ocean viral communities. *Science* 348. DOI:
585   10.1126/science.1261498.

586 Brum JR, Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine
587   virology. *Nature Reviews Microbiology* 13. DOI: 10.1038/nrmicro3404.

588 Bushnell B. 2015.BBMap

589 Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, Behrenfeld MJ, Boetius A, Boyd
590   PW, Classen AT, Crowther TW, Danovaro R, Foreman CM, Huisman J, Hutchins DA, Jansson JK,
591   Karl DM, Koskella B, Mark Welch DB, Martiny JBH, Moran M, Orphan V, Reay D, Remais JV,
592   Rich V, Singh BK, Stein LY, Stewart FJ, Sullivan MB, van Oppen MJH, Weaver SC, Webb EA,
593   Webster N. 2019. Scientists' warning to humanity: microorganisms and climate change. *Nature
594   Review Microbiology* In Press:1–9. DOI: .1037//0033-2909.I26.1.78.

595 Cook R, Hooton S, Trivedi U, King L, Dodd CER, Hobman JL, Stekel DJ, Jones MA, Millard AD. 2020.
596   Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the
597   potential to alter the metabolism and virulence of veterinary pathogens. *bioRxiv*:2020.10.08.329714.
598   DOI: 10.1101/2020.10.08.329714.

599 De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: Visualizing and
600   processing long-read sequencing data. *Bioinformatics* 34:2666–2669. DOI:
601   10.1093/bioinformatics/bty149.

602 Danecek P, McCarthy SA. 2017. BCFtools/csq: Haplotype-aware variant consequences. *Bioinformatics*.

603   DOI: 10.1093/bioinformatics/btx100.

604   Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB. 2014.
605   Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature*
606   513:242–245. DOI: 10.1038/nature13459.

607   Duhaime MB, Deng L, Poulos BT, Sullivan MB. 2012. Towards quantitative metagenomics of wild
608   viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of
609   the linker amplification method. *Environ Microbiol* 14:2526–2537. DOI: 10.1111/j.1462-
610   2920.2012.02791.x.

611   Duhaime MB, Solonenko N, Roux S, Verberkmoes NC, Wichels A, Sullivan MB. 2017. Comparative
612   Omics and Trait Analyses of Marine Pseudoalteromonas Phages Advance the Phage OTU Concept
613   . *Frontiers in Microbiology* 8:1241.

614   Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE,
615   Boyd JA, Hodgkins SB, Wilson RM, Trubl G, Li C, Frolking S, Pope PB, Wrighton KC, Crill PM,
616   Chanton JP, Saleska SR, Tyson GW, Rich VI, Sullivan MB. 2018. Host-linked soil viral ecology
617   along a permafrost thaw gradient. *Nat Microbiol* 3:870–880. DOI: 10.1038/s41564-018-0190-y.

618   Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova
619   K, Carmichael M, Cruaud C, Dimier C, Domínguez-Huerta G, Ferland J, Kandels S, Liu Y, Marec
620   C, Pesant S, Picheral M, Pisarev S, Poulain J, Tremblay JÉ, Vik D, Acinas SG, Babin M, Bork P,
621   Boss E, Bowler C, Cochrane G, de Vargas C, Follows M, Gorsky G, Grimsley N, Guidi L, Hingamp
622   P, Iudicone D, Jaillon O, Kandels-Lewis S, Karp-Boss L, Karsenti E, Not F, Ogata H, Poulton N,
623   Raes J, Sardet C, Speich S, Stemmann L, Sullivan MB, Sunagawa S, Wincker P, Culley AI, Dutilh
624   BE, Roux S. 2019. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* 177:1109-
625   1123.e14. DOI: 10.1016/j.cell.2019.03.040.

626   Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome
627   assemblies. *Bioinformatics* 29:1072–1075. DOI: 10.1093/bioinformatics/btt086.

628   Hurwitz BL, Deng L, Poulos BT, Sullivan MB. 2013. Evaluation of methods to concentrate and purify
629   ocean virus communities through comparative, replicated metagenomics. *Environ Microbiol*
630   15:1428–1440. DOI: 10.1111/j.1462-2920.2012.02836.x.

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

634   Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat
635   graphs. *Nature Biotechnology* 37:540–546. DOI: 10.1038/s41587-019-0072-8.

636   Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: Scalable and
637   accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*
638   27:722–736. DOI: 10.1101/gr.215087.116.

639   Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
640   *preprint*:1303.3997. DOI: arXiv:1303.3997 [q-bio.GN].

641   Li H. 2016. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences.

642   *Bioinformatics*. DOI: 10.1093/bioinformatics/btw152.

643   Li H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*. DOI:
644   10.1093/bioinformatics/bty191.

645   Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The
646   Sequence Alignment/Map format and SAMtools. *Bioinformatics*. DOI:
647   10.1093/bioinformatics/btp352.

648   De Maio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, AbuOun M, Stubberfield
649   E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Walker
650   AS, Stoesser N, consortium on behalf of the R. 2019. Comparison of long-read sequencing
651   technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics* 5:e000294.
652   DOI: 10.1099/mgen.0.000294.

653   Marijon P, Chikhi R, Varré J-S. 2020. yacrd and fpa: upstream tools for long-read genome assembly.
654   *Bioinformatics*. DOI: 10.1093/bioinformatics/btaa262.

655   Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, De La Cruz Peña MJ, Martínez JM,
656   Anton J, Gasol JM, Rosselli R, Rodriguez-Valera F, Sullivan MB, Acinas SG, Martinez-Garcia M.
657   2017. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature*
658   *Communications* 8. DOI: 10.1038/ncomms15892.

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

661   Moss EL, Maghini DG, Bhatt AS. 2020. Complete, closed bacterial genomes from microbiomes using
662   nanopore sequencing. *Nature Biotechnology* 38:701–707. DOI: 10.1038/s41587-020-0422-6.

663   Nayfach S, Camargo AP, Eloe-fadrosh E, Roux S, Kyrpides N. 2020. CheckV: assessing the quality of
664   metagenome-assembled viral genomes. *bioRxiv*:1–20.

665   Nelson WC, Maezato Y, Wu YW, Romine MF, Lindemann SR. 2016. Identification and resolution of
666   microdiversity through metagenomic sequencing of parallel consortia. *Applied and Environmental*
667   *Microbiology*. DOI: 10.1128/AEM.02274-15.

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

670   Oxford Nanopore Technologies Ltd. 2018. Medaka.

671   Roux S. 2015. ClusterGenomes.

672   Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E,
673   Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A,
674   Duarte CM, Gasol JM, Vaqué D, Bork P, Acinas SG, Wincker P, Sullivan MB, Vaque D, Bork P,
675   Acinas SG, Wincker P, Sullivan MB, Coordinators TO. 2016. Ecogenomics and potential
676   biogeochemical impacts of globally abundant ocean viruses. *Nature* 537:689–693. DOI:
677   10.1038/nature19366.

678   Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. 2017. Benchmarking viromics: an in silico

679   evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*
680   5:e3817. DOI: 10.7717/peerj.3817.

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

683   Shagin DA, Lukyanov KA, Vagner LL, Matz M V. 1999. Regulation of average length of complex PCR
684   product. *Nucleic acids research* 27:e23-i–e23-iii. DOI: 10.1093/nar/27.18.e23-i.

685   Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang
686   W, Xu J, Holmes EC, Zhang YZ. 2016. Redefining the invertebrate RNA virosphere. *Nature*. DOI:
687   10.1038/nature20167.

688   Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, McDonnell SA, Khokhlova E
689   V., Draper LA, Forde A, Guerin E, Velayudhan V, Ross RP, Hill C. 2019. The Human Gut Virome
690   Is Highly Diverse, Stable, and Individual Specific. *Cell Host and Microbe* 26:527-541.e5. DOI:
691   10.1016/j.chom.2019.09.009.

692   Trubl G, Jang H Bin, Roux S, Emerson JB, Solonenko N, Vik DR, Solden L, Ellenbogen J, Runyon AT,
693   Bolduc B, Woodcroft BJ, Saleska SR, Tyson GW, Wrighton KC, Sullivan MB, Rich VI. 2018. Soil
694   Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems* 3:1–21. DOI:
695   10.1128/msystems.00076-18.

696   Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long
697   uncorrected reads. *Genome Research* 27:737–746. DOI: 10.1101/gr.214270.116.

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

701   Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB,
702   Temperton B. 2019. Long-read viral metagenomics captures abundant and microdiverse viral
703   populations and their niche-defining genomic islands. *PeerJ* 7:e6800. DOI: 10.7717/peerj.6800.

704   Wick RR. 2017. Porechop. *Github https://github. com/rrwick/Porechop*.

705   Wick RR, Holt KE. 2020. Benchmarking of long-read assemblers for prokaryote whole genome
706   sequencing. *F1000Research* 8. DOI: 10.12688/f1000research.21782.3.

707   Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, Krupovic M, Fire A, Dolja V V, Koonin E V.
708   2020. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome.
709   *Nature Microbiology* 5:1262–1270. DOI: 10.1038/s41564-020-0755-4.

710   Yuan Y, Gao M. 2017. Jumbo Bacteriophages: An Overview. *Frontiers in microbiology* 8:403. DOI:
711   10.3389/fmicb.2017.00403.

712   Zhong Z-P, Rapp JZ, Wainaina JM, Solonenko NE, Maughan H, Carpenter SD, Cooper ZS, Jang H Bin,
713   Bolduc B, Deming JW, Sullivan MB. 2020. Viral Ecogenomics of Arctic Cryopeg Brine and Sea
714   Ice. *mSystems*. DOI: 10.1128/msystems.00246-20.

715

716

717

718

719

720

Figure 1

Overview of wet lab optimization experiments and informatic benchmarking.

(A) Laboratory optimization ('Experiment 1') in which a mock community of three phages was used to conduct three experiments aimed at producing longer reads from less input DNA. (B) Informatics benchmarking ('Experiment 2') in which a seawater virome was sequenced with short-reads (Illumina) and long-read sequencing. Three distinct long-read libraries were generated, error-corrected and assembled, and were compared to short-read assemblies to assess accuracy and assembly performance.
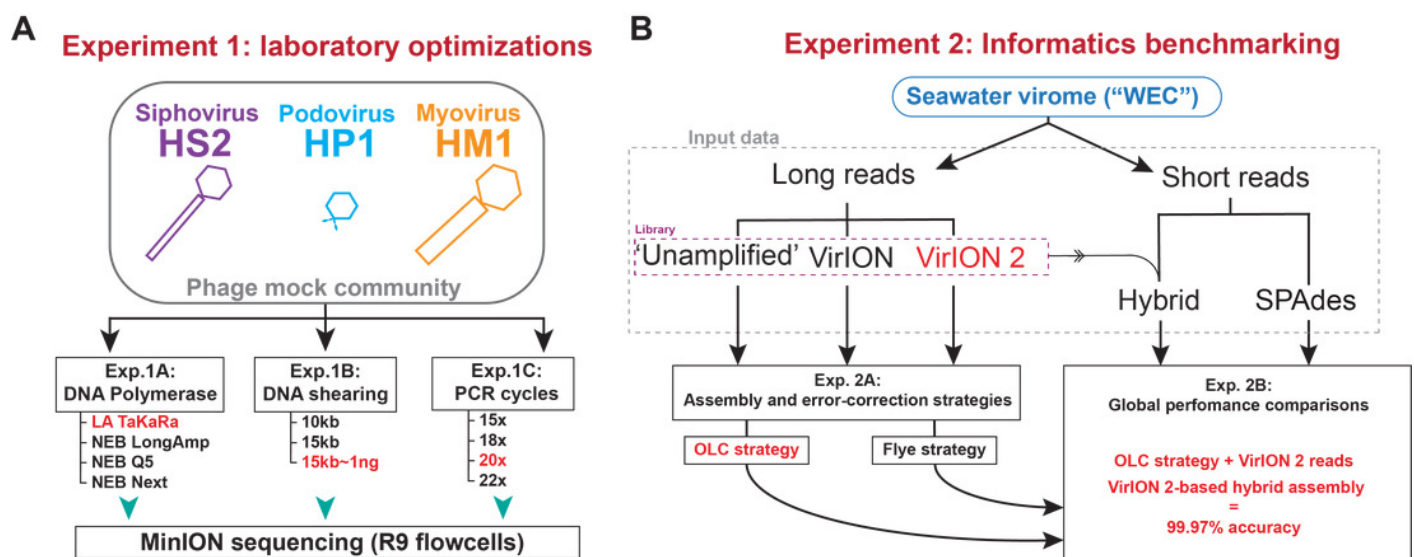
Figure 2

Laboratory optimization yield longer reads from less DNA.

(A) Boxplots showing the median and quartiles of the read length distribution between four DNA polymerases. (B) Boxplots showing the median and quartiles of the read length distribution between DNA shearing size treatments and one low-input DNA (1ng) variant of the 15kbp shearing treatment. (C) Boxplots showing the median and quartiles of the read length distribution of three long-read library types, either unamplified or amplified (VirION and VirION2). (D) Boxplots showing the median and quartiles of the read length distribution between four thermocycling treatments (here, number of cycles). Asterisks represent a significant difference ($p<0.0001$) between pairs of replicate treatments where applicable.
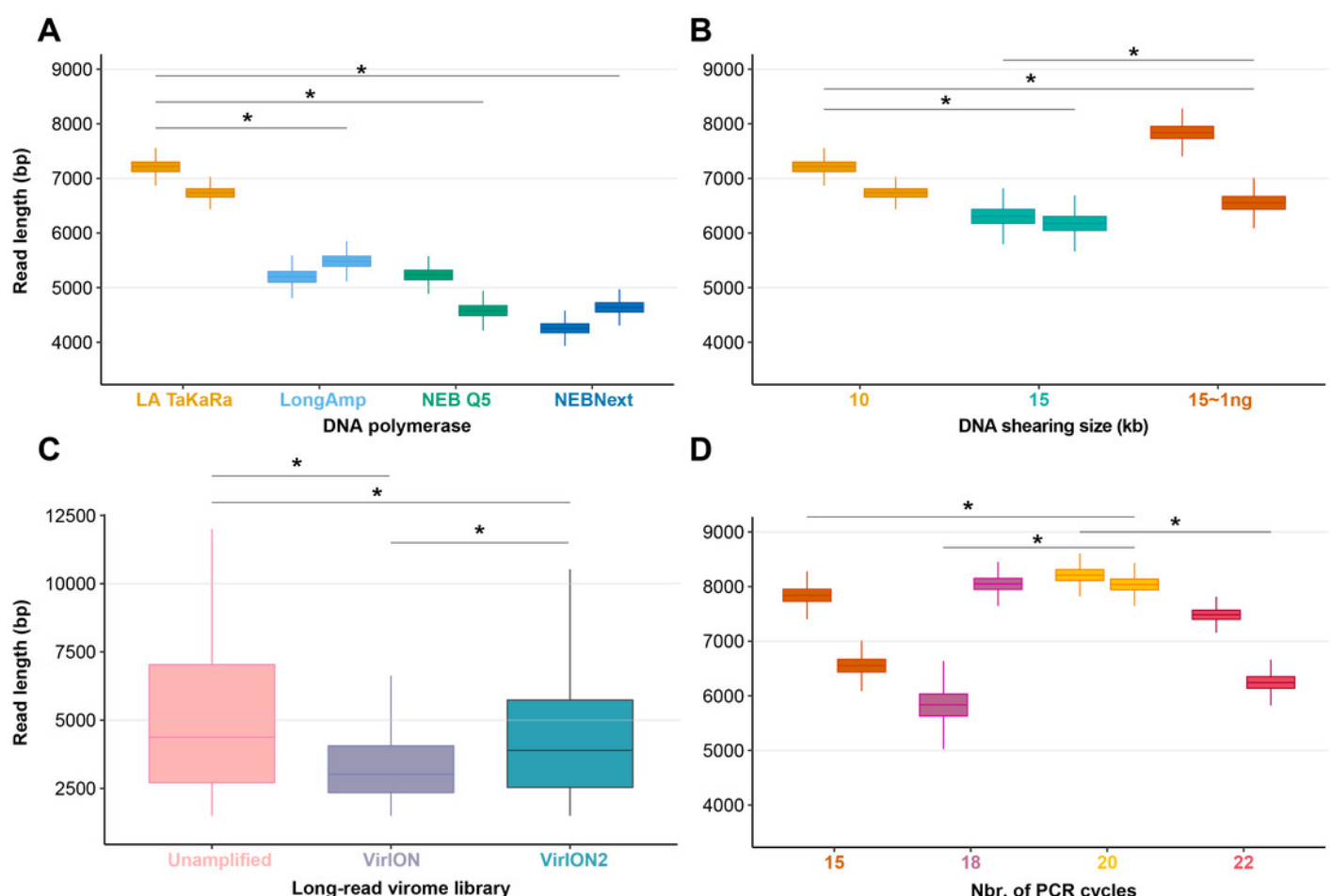
Figure 3

Error-correction profiles between library methods and assembly strategies using the WEC sample.

(A) On the x-axis, mismatches and insertion/deletion ('indels') events according to assembly strategy (full OLC, full Flye, and Hybrid) are grouped separately, and divided into three facets, one for each long-read library method (unamplified, VirION, VirION2). The number of errors (y-axis) is scaled to the binary logarithm (log2) for scale fitting purposes. (B) Boxplot depicting the protein size distribution (in amino acids, denoted as 'a.a'; y-axis) derived from each library method (x-axis), each of which is sub-grouped per assembly strategy. In both A and B panels, there were no results from Flye assemblies for the raw datasets, as these could not be produced.
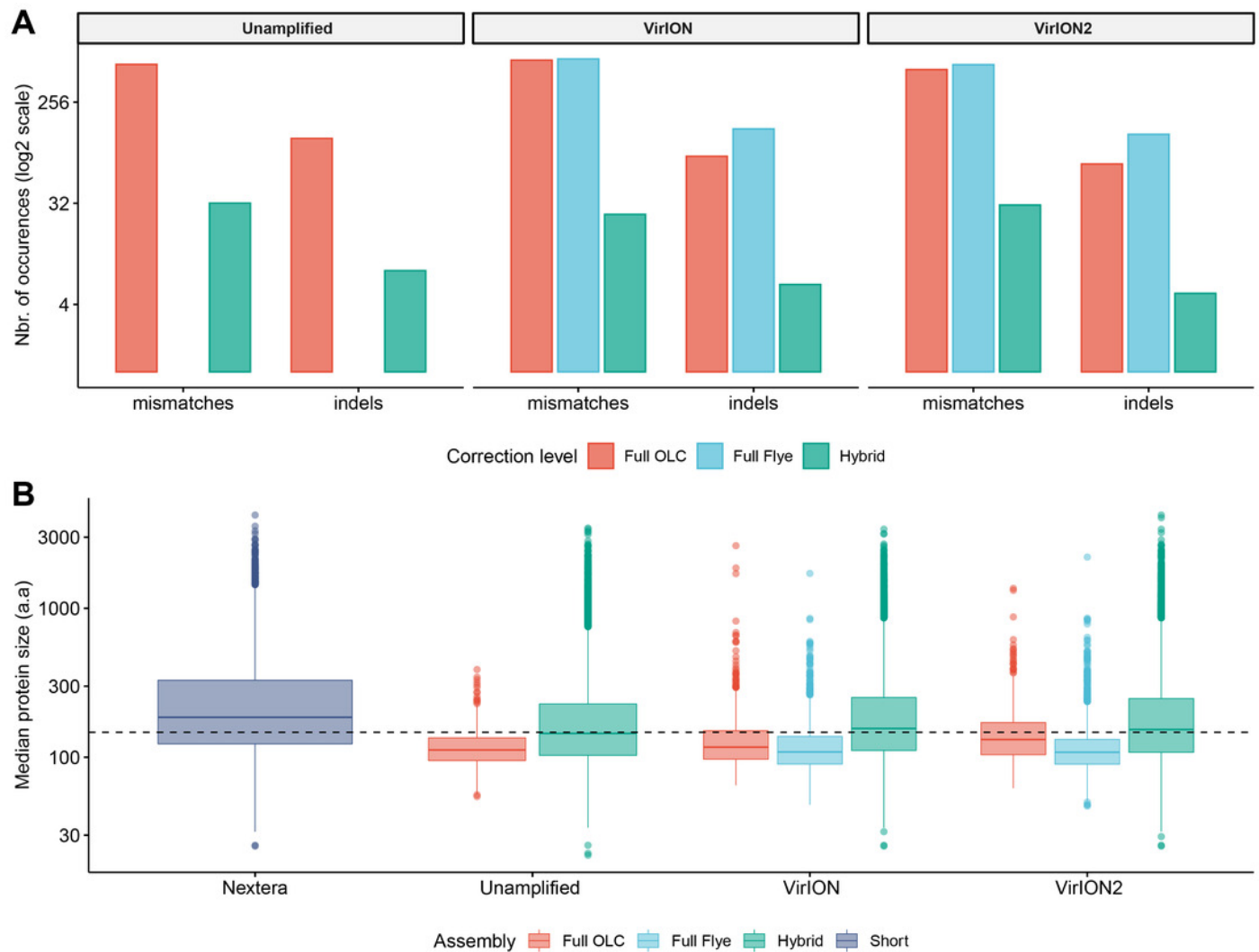
Figure 4

Comparison of virus genome properties between short-read and 'long-read-enhanced' viromes.

(A) Workflow to produce 'enhanced viromes', in which Spades, hybrid and long-read (OLC) viruses are combined to maximize the recovery of virus signals. (B) Cumulative Distribution Function (CDF) plot depicting the frequency (y-axis) of virus genomes according to genome length (measured in kilo basepairs (kbp), x-axis) between three assembly strategies. (C) Cumulative Distribution Function (CDF) plot depicting the frequency (y-axis) of virus genomes according to genome 'completeness' (measured in %, x-axis) between three assembly strategies. (D) Cumulative Distribution Function (CDF) plot depicting the frequency (y-axis) of virus genomes according to genome microdiversity per genome (measured as π, x-axis) between three assembly strategies.
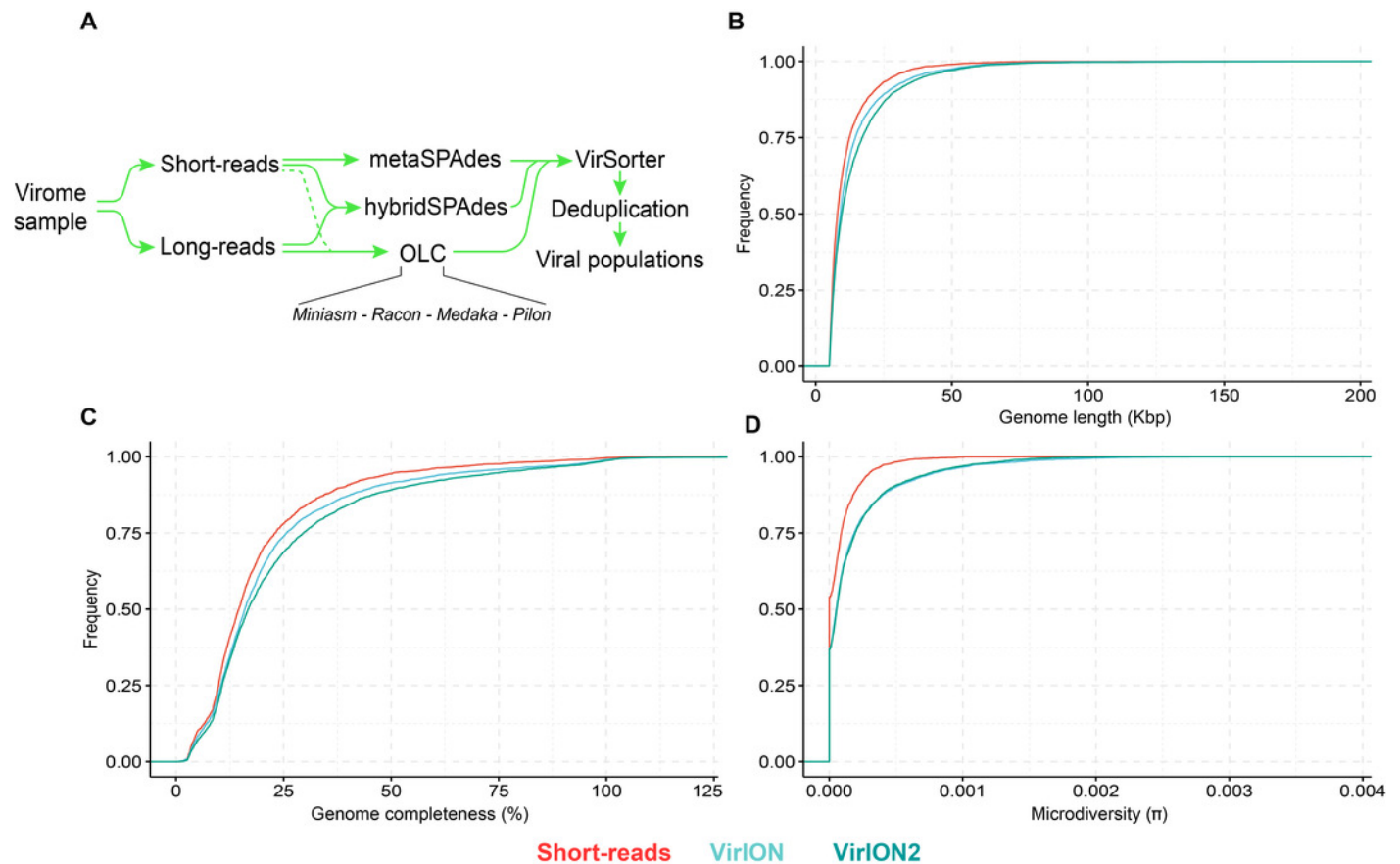
Figure 5

Additional insights are gained through a VirION2-enhanced assembly strategy.

(A) Rank abundance curve depicting the seawater virus community, colored according to whether a virus population (individual bars) was detected uniquely (turquoise) or in multiple (pastel red) assembly types. The top 10% most abundant viral populations are highlighted between dashed lines, where they are divided per assembly origin in the pie chart. (B) Boxplots depicting the level of microdiversity between shared and unique viral populations within each constituent assembly present in the 'enhanced' dataset.