Genome-resolved year-round dynamics reveal a broad range of giant virus microdiversity

Yue Fang[1,•], Lingjie Meng[1,•], Jun Xia[1], Yasushiro Gotoh[2], Tetsuya Hayashi[2], Keizo Nagasaki[3], Hisashi Endo[1], Yusuke Okazaki[1], Hiroyuki Ogata[1,*]

1. Institute for Chemical Research, Kyoto University, Gokasho, Uji, Japan
2. Department of Bacteriology, Faculty of Medical Sciences, Kyushu University, Maidashi, Higashi-ku, Fukuoka, Japan
3. Faculty of Science and Technology, Kochi University, Nankoku, Kochi, Japan

•These authors equally contributed to this work.

Corresponding author

Hiroyuki Ogata (Tel: +81-774-38-3270; E-mail: ogata@kuicr.kyoto-u.ac.jp)

Abstract

Giant viruses are crucial for marine ecosystem dynamics because they regulate microeukaryotic community structure, accelerate carbon and nutrient cycles, and drive the evolution of their hosts through co-evolutionary processes. Previously reported long-term observations revealed that these viruses display fluctuations in abundance. However, the underlying genetic mechanisms driving such dynamics in these viruses remain largely unknown. In this study, we investigated population and intra-population dynamics of giant viruses using time-series metagenomes from eutrophic coastal seawater samples collected over 20 months. A newly developed near-automatic computational pipeline generated 1,065 high-quality metagenome-assembled genomes covering six major giant virus lineages. These genomic data revealed year-round recovery of the viral community at the study site and distinct dynamics of different viral populations classified as persistent (n = 9), seasonal (n = 389), sporadic (n = 318), or others. Notably, year-round recovery patterns were observed at the intra-population genetic diversity level for viruses classified as persistent or seasonal. Our results further indicated that the viral genome dynamics were associated with intra-population diversity; specifically, giant viruses with broader niche breadth tended to exhibit greater levels of microdiversity. We argue that greater microdiversity in viruses likely enhances adaptability and thus survival under the virus–host arms race during prolonged interactions with their hosts.

Introduction

Microdiversity refers to subspecies-level (intra-population) genomic diversity [1–3]. It can alter physiological characteristics[4], differentiate ecological niches[5], and maintain the stability of microbial populations[6]. The driving forces that generate microdiversity include genetic factors, such as mutations, horizontal gene transfer, and genomic rearrangements, and selective pressures, such as temperature, light, predator, and nutrition. Microdiversity is shaped by the complex interplay of these factors and therefore provides a framework for understanding the eco-evolutionary dynamics of the microbial world. Despite its importance, research focused on microdiversity in viruses has been less common than studies on their macrodiversity. Additionally, an increasing body of literature has underscored the importance of subspecies variation in viruses. A few mutations in a single protein, such as the spike protein in coronaviruses, can significantly alter viral infectivity and, consequently, epidemicity[7]. Similarly, one single mutation expands the host range of *Pseudomonas* phage LUZ7[8]. Therefore, understanding virus microdiversity is crucial for enhancing our knowledge on their impact in nature.

Viruses play a critical role in marine ecosystems by modulating microbial community composition and participating in biogeochemical cycles[9,10]. Members of the viral phylum *Nucleocytoviricota*, often referred to as giant viruses, are widespread[11], abundant[12], and active[13] in the ocean. Some of these viruses contribute to the global carbon export by infecting plankton[14]. Temporal dynamics are known to reflect the interactions of giant viruses with their hosts and the population structure under environmental pressures. *Emiliania huxleyi* viruses tend to retain the same genotypes throughout the bloom periods of their hosts[15]. Generally, the *Imitervirales* community exhibits synchronous seasonal cycles with eukaryotes and year-round recurrence; nonetheless, most individual viral populations tend to be specialists rather than generalists[16]. To determine what underlying forces are driving viral niche differentiation, a thorough and comprehensive investigation is needed into the structure, dynamics, and diversity of viruses at the subspecies level.

Profiling giant viral microdiversity and dynamics has numerous challenges. Previous studies predominantly employed marker-based approaches, which do not capture variability for all populations[17] nor offer genome-wide evidence of selection[18]. Therefore, high-quality and fine-resolution genomes are necessary for comprehensive investigation. Additionally, the scarcity of time-series data limits the ability to track giant virus population dynamics over time. Moreover, achieving adequate sequencing depth is also necessary for assembling and capturing signals of microdiversity,

64 especially because the lower abundance of giant viruses makes them harder to detect compared with
65 the overwhelming signals of prokaryotes and phages[9,12].

66 Here, we analyzed 42 coastal samples collected during 20 months from January 2017 to
67 September 2018. The time-series samples were subjected to deep sequencing, which yielded over 14
68 billion metagenomic reads. We also developed a nearly automatic pipeline for generating
69 metagenome-assembled genomes (MAGs) of giant viruses; this pipeline consisted of assembly,
70 binning, screening, quality control, deduplication, and quality assessment. Applying this pipeline to
71 42 coastal metagenomes resulted in the creation of a coastal giant virus genome database containing
72 1,065 non-redundant giant virus MAGs that serve as representative species-level genomes.
73 Furthermore, we profiled the intra-species nucleotide-resolved microdiversity of giant viruses
74 through metagenomic read mapping and revealed their population structure over 20 months across
75 two years. Collectively, the results from these two methods provided fine-scale insights into the
76 ecological roles and evolutionary trajectory of giant viruses.
77

78 # Methods

79 ## Sample Collection, DNA Extraction, and Sequencing

80 The water samples used in this study were same as the ones in a previous study[16] (Table S1).
81 Briefly, seawater samples were collected from three adjacent sites, labeled as 'J' (33°25'43.2"N
82 133°22'49.5"E), 'M' (33°25'60.0"N 133°24'38.3"E), and 'F' (33°26'33.6"N 133°24'41.8"E), in
83 Uranouchi Inlet, Kochi Prefecture, Japan (Fig. S1a). In total, 42 samples were obtained over 20
84 months from 5 January 2017 to 25 September 2018 (Fig. S1b). Because the sampling strategy
85 focused on capturing bloom events, a higher sampling frequency was performed during the summer
86 months. Seawater samples (10 L for each sample) were collected from a depth of 5 m and
87 subsequently transported to the laboratory for filtration. The samples were sequentially filtered
88 through 3.0-μm and 0.8-μm filters (diameter 142 mm, Merck, Darmstadt, Germany), followed by a
89 0.22-μm filter through a Sterivex filtration unit (Merck, Darmstadt, Germany). After filtration, the
90 filters were stored at −80°C until DNA extraction.

91 DNA extraction was performed in October 2020 using an in-house protocol[19]. Frozen 0.22-
92 μm filters, which included organisms and viruses collected in 0.22–0.8-μm size fractions, were
93 transferred to 1.5-mL microtubes containing 0.1-mm glass beads (0.2 g) and then filled with
94 xanthogenate buffer (1 M Tris-HCl, 0.5 M ethylenediaminetetraacetic acid [EDTA], 5 M ammonium
95 acetate, 10% potassium xanthogenate, 10% SDS, sterile water). Bead beating (Taitech, Beavercreek,

OH, USA) was used to lyse cells and virions, followed by a 60-minute incubation at 70°C to increase DNA yield. Glass beads were removed from the mixture after centrifugation. Then, 600 µL isopropanol was added to the supernatant and mixed. The precipitated DNA was purified with a NucleoSpin gDNA Clean-up Kit (Macherey-Nagel, Düren, Germany) and then dissolved in a Tris-ethylenediaminetetraacetic acid (Tris-EDTA) buffer. A Qubit 4 Fluorometer (Invitrogen, Carlsbad, CA, USA) and the 4150 TapeStation system (Agilent, Santa Clara, CA, USA) were used to measure the quantity and quality of yielded DNA. Extracted DNA were stored at −20°C until sequencing.

Of the 42 extracted DNA samples, 27 were sequenced by Kyushu University (Fukuoka, Japan) and 15 were sequenced by Macrogen Japan (Tokyo, Japan). Both facilities used the same sequencing kits and platforms. The sequencing library was prepared using TruSeq Nano DNA Kit (Illumina, San Diego, CA, USA; insert size, 350 bp), and Illumina NovaSeq 6000 was used to sequence the DNA in paired-end mode (2 × 151 bp, 45 Gbp data for each sample). Quality control of raw sequencing reads was conducted using FastQC v0.11.8[20], followed by trimming with Trimmomatic v0.39[21] for adaptors and low-quality reads and fastp v0.20.1[22], which specifically addressed poly-G tail issues in NovaSeq sequencing. 42 metagenomes are available at the DNA Data Bank of Japan (DDBJ) under the BioProject Accession PRJDB12848.

Generation of Giant Virus MAGs

The overall workflow for generating giant virus MAGs is outlined in a schematic diagram (Fig. S2). Briefly, the pipeline comprises the following steps:

1) Individual metagenome assembly and binning

For each metagenome, trimmed reads were assembled into contigs using MEGAHIT v1.2.9[23] in 'meta-large' mode. Bowtie2 v2.4.5[24] mapped reads to assembled contigs longer than 1 Kbp, which were then sorted and converted to BAM files using samtools v1.16[25]. The jgi_summarise_bam_contig_depths script of MetaBAT2[26] was used to calculate average mapping depth for each contig. Contigs were clustered into bins using MetaBAT2 v2.12.1[26], a binning tool that uses both coverage depth and normalized tetra-nucleotide frequency information. Contigs shorter than 2500 bp were excluded from the binning, and the default MetaBAT2 parameters were employed. For each of the 42 metagenomes, a corresponding set of bins was generated (Fig. S2a).

2) Giant virus screening

Previous genome-resolved metagenomic studies for giant viruses have used a fixed set of markers or a single marker gene to detect giant viruses[27–29]. However, these methods are limited by the incompleteness of giant virus MAGs, which may result in overlooking some giant virus genomes or in false-positive detection. A correlation pattern was previously observed between genome size and number of core genes[30]. Therefore, we used a core gene density index to screen potential giant viruses. Briefly, we selected 20 marker genes of Nucleo-Cytoplasmic Virus Orthologous Groups (NCVOGs), which are universally or nearly universally present across known giant virus families[31]. Then, we assigned weights to 20 NCVOGs according to conservation levels in individual families (from 0 for absence to 1 for the conservation across lineages). Then, the index was calculated using the following equation:

$$Density\ index = \frac{\sum_{k=1}^{20} weight_k}{log10(genome\ size) - 4}$$

To assess the effectiveness of this metric, we compiled a genome database comprising 205 reference giant virus genomes and 6,497 cellular genomes ($N_{archaea}$ = 334, $N_{bacteria}$ =6,114, $N_{eukaryota}$ = 49) downloaded from the KEGG database in June 2019. This density index effectively distinguished giant virus genomes from those of cellular organism genomes (Fig. S2b). Furthermore, our study revealed a distinct gap in the indices of all raw bins created by MetaBAT2, where the gap (corresponding to density index = 5.75) was the same as the one that discriminates reference viral genomes from cellular genomes (Fig. S2c and Supplementary Methods for more detail). Therefore, we used the core gene density index to screen for potential giant viruses.

3) Refinement of giant virus MAGs

To improve the quality of the preliminarily screened giant virus MAGs, we employed a rigorous quality control pipeline comprised of four principal steps (Fig. S2d): (1) verification of giant virus-specific traits in contigs, (2) removal of non-giant virus MAGs from the screened collection, (3) decontamination of viral MAGs to eliminate non-viral sequences, and (4) resolution of chimeric bins to separate mixed lineage bins. Detailed descriptions of these methodologies are provided in the Supplementary Methods. A visual summary of the quality control procedures is depicted in the schematic representation (Fig. S2d).

4) Deduplication of MAGs

To construct a representative set of genomes, we used the dereplication tool dRep v3.2.2[32] to remove redundant MAGs (parameter: --S_algorithm ANImf --ignoreGenomeQuality -pa 0.9 -sa 0.95) (Fig. S2e). dRep performs clustering of MAGs by comparing pairwise average nucleotide identity (ANI), using Mash for initial rapid screening with a minimum ANI threshold of 90% to group closely related genomes. For more refined and accurate clustering, ANImf was applied with a higher minimum ANI threshold of 95%. The threshold was established based on the distribution of ANI values across all MAGs (Fig. S3), which showed that an 95% ANI value represented a practical boundary for population delineation. In addition, the 95% ANI value has been used as the demarcation criteria to define the viral species for medusaviruses[33] and imiterviruses[34]. With the parameter '--ignoreGenomeQuality,' dRep assigned scores to MAGs based on N50, genome size, and centrality. The highest-scoring MAG in each cluster was selected as the representative genome of a giant virus population. Finally, 1,065 giant virus MAGs were obtained after deduplication and used as representative species-level giant virus genomes.

An assessment of the performance of this pipeline is included in the supplementary information.

Taxonomic Classification

Taxonomic classification was determined by referring to a phylogenetic tree that incorporated the 1,065 generated representative species-level MAGs and 205 giant virus reference genomes. This tree was based on a concatenated alignment of the three hallmark genes in the viral informational module (RNApolA, RNApolB, DNApolB), which comprehensively reveal virus evolution. The hallmark genes were detected using a python program, 'ncldv_markersearch'[28]. Multiple sequence alignments of the hallmark genes were performed using MAFFT v7.505[35], followed by concatenation of the three genes. Then, unconserved positions were removed using TrimAl v1.4.1[36] with gap threshold of 0.1. A phylogenetic tree was constructed using IQ-TREE v2.2.0[37] with the LG+F+I+G4 model, which was recommended in a previous study[38]. For *Mirusviricota*, an additional HK97-fold major capsid protein tree was reconstructed.

Phylogeny-Informed MAG Assessment (PIMA)

In this study, we developed an approach, phylogeny-informed MAG assessment (PIMA), to assess the quality of giant virus MAGs. This approach was designed to overcome the limitations caused by a lack of reference genomes, which impacts the accuracy of quality assessments for giant

189 virus MAGs. This approach requires a guide tree of giant viruses that has been rerooted in
190 accordance with the latest taxonomic classifications and evolutionary scenarios[38,39]. Then, the
191 relative evolutionary divergence (RED) values were calculated to classify taxonomic levels for each
192 clade (e.g., order and family)[38]. Within a specific clade, MAG genes were annotated with orthologous
193 groups (OGs); then, core genes in this clade were defined as those identified in more than 50% of
194 the genomes in the clade. We assessed MAG consistency and redundancy using the following
195 equations:

$$Consistency = \frac{number\ of\ core\ genes\ in\ a\ MAG}{number\ of\ core\ genes\ in\ the\ clade}$$

$$Redundancy = \frac{number\ of\ redundant\ genes\ in\ a\ MAG}{number\ of\ core\ genes\ in\ a\ MAG}$$

196 Redundant genes in a MAG are defined as genes with more copies than the mode copy
197 number (the most common number of copies) for the given genes across all MAGs in the evaluated
198 lineage.
199 The details of these methods are included in the Supplementary Methods. In our study, we
200 adopted a RED value of 0.65 as the threshold for clade definition, which corresponded to the level
201 of viral genus or family[38]. MAG quality within each clade was assessed using the above equations
202 within the clade. The consistency and redundancy distribution of the MAGs is shown in Fig. S4a.
203
204 ## Population and Community Dynamics
205 To assess the time-series dynamics of giant viruses, metagenomic reads were back-mapped to
206 1,065 representative MAGs using Bowtie2 v2.4.5[24]. Reads per kilobase per million (RPKM) was
207 employed to normalize and profile abundance of giant virus MAGs using CoverM v0.4.0[40]. MAGs
208 were defined being present in a sample only if they exhibited a coverage breadth (proportion of
209 genomes with at least one read mapped) exceeding 50% of the whole MAG size. Then, we defined
210 'abundant' giant viruses in one sample based on the contribution of the viral MAG to the viral
211 community diversity of the sample, measured by the Simpson index[41]. For each sample, the initial
212 step involved calculating the overall Simpson diversity index across all giant virus MAGs.
213 Subsequently, these MAGs were sorted by their relative abundance (RPKM) in descending order.
214 Viral MAGs were considered abundant if, once cumulated, they represented the top 80% of the
215 whole sample diversity based on the Simpson index values.

To assess the time-series dynamics of giant viruses at the community level, the Sorensen–Dice dissimilarity measure was employed to compare community compositions at different times using the vegdist function in the R package 'vegan'[42] (method = 'bray,' binary = T). Because of the uneven frequency of sampling across months, we binned the 42 metagenomes into monthly intervals (N = 20) and calculated the averages of all variables for the samples within the same month. This strategy was supported by the observation that samples from the same month exhibited the most similar compositions.

Genetic Diversity of Viral Populations

The back-mapping files were used as input for InStrain v1.0.0[43], which facilitated the generation of single nucleotide variant (SNV) profiles on both the genome and gene scales. SNVs were exclusively identified at sites with a minimum coverage depth of 5×. The nucleotide diversity (ND), an index of genetic diversity within a population, was calculated according to the method described by Nei and Li in 1979[44]:

$$1 - (f_A^2 + f_C^2 + f_G^2 + f_T^2)$$

In the equation, fx refers to the frequency of the nucleotide X (A, C, G, or T) at a given nucleotide site. The genome-wide ND value was calculated as the average across all SNV sites within a genome.

In addition to ND, SNV/Mb (the number of SNV sites per million base pairs) was used as an additional metric for assessing microdiversity. The microdiversity analyses focused on abundant viral MAGs because few reads are not informative and microdiversity metrics for rare MAGs could be unreliable.

Definition of Niche Categories

Each giant virus MAG that had appeared before the second winter (December 2017) was classified into one of three niche categories based on its occurrence pattern: persistent, seasonal, or sporadic. The classification criteria for each MAG were based on their occurrence pattern within 8-month sliding windows (1-month step size). The first window spanned January 2017 to August 2017, and the final window covered February 2018 to September 2018. We excluded MAGs that were exclusively present in 2018 and absent in 2017. The definitions of the three niche categories were as follows:

1) Persistent: A MAG was classified as 'persistent' if it appeared in more than four consecutive months within at least one of the 8-month windows.

2) Seasonal: A MAG that was not persistent was classified as 'seasonal' if its appearance in 2017 was exclusively in the months that could be entirely covered by at least one 8-month window of 2017, reappeared in the same monthly window of 2018, and did not appear in the remaining months outside that window in 2018. For example, a MAG that appeared in January and September in 2017 was excluded from being classified as seasonal.

3) Sporadic and Other: A MAG that did not follow the patterns of persistent or seasonal was classified as 'sporadic' or 'other.' Specifically, if the MAG only appeared in 2017, it was labeled as sporadic. If a MAG appeared in both 2017 and 2018, it was classified as other (N = 28).

Additionally, the Levins' index, which represents niche breadth (B) of a giant virus[16], for each giant virus was calculated by the 'spaa' package in R, which used the following formula:

$$B_j = \frac{1}{\sum_i^n p_i^2}$$

In the equation, p_i is the proportion of virus j in month i relative to all months.

Fixation Index

The fixation index is a measure of population differentiation due to genetic structure. It is frequently estimated from single nucleotide polymorphisms. To estimate the fixation index, we followed a previous pipeline[45] and first calculated the distance (*pi*) between two samples for a given MAG at an SNV site as follows:

$$pi(a,b) = a_A\left(\sum_{i=1}^{3} b_{non\text{-}A}\right) + a_C\left(\sum_{i=1}^{3} b_{non\text{-}C}\right) + a_T\left(\sum_{i=1}^{3} b_{non\text{-}T}\right) + a_G\left(\sum_{i=1}^{3} b_{non\text{-}G}\right)$$

Here, *a* and *b* are vectors representing the nucleotide frequencies of one SNV site of a genome in samples *a* and *b*, respectively. For example, if there are 30 'A,' 0 'C,' 10 'T,' and 20 'G' mapped at the same site, the vector for this site will be (0.5, 0, 0.33, 0.66).

Then, the fixation index was calculated using the average distance *pi* with the following equation. A genome-wide *pi* value was calculated as the average of the *pi* values of all SNV sites shared by two genomes:

$$fst(a,b) = 1 - \frac{\left(\frac{\overline{pi}(a,a) + \overline{pi}(b,b)}{2}\right)}{\overline{pi}(a,b)}$$

Genetic structure similarity was measured by $1 - fst$.

Statistical Analysis

The Wilcoxon rank-sum test was employed to identify significant differences in nucleotide diversity, coverage, and SNV/Mb between categorical groups. P-values were corrected using the Benjamini–Hochberg procedure in R, and adjusted P-values <0.05 were considered significant. Visualization was done using R package 'ggplot2'[46], DiGAlign v2.0[47], Cytoscape v3.7.1[48] and iTol v6[49].

Results

Japan Coastal Giant Virus Genomes

A total of 2,655,994 contigs were assembled and 16,110 raw bins were generated from 42 coastal metagenomes extracted from water samples of 0.22–0.8-μm size fractions in Uranouchi Inlet, Kochi Prefecture, Japan (Fig. S1; Table S1). Then, we screened 3,082 potential giant virus bins (19.13% of all raw bins; range, 25–147 from individual metagenomes), subsequently refined them to enhance their quality, and retained 2,635 giant virus MAGs (16.4% of the total bins). The pairwise ANI of these giant virus MAGs showed a bimodal distribution akin to that observed in prokaryotes (Fig. S3)[50], which indicated that the boundary for populations of coastal giant viruses was around 98% ANI; this was consistent with findings from a previous study[29]. As a result, a collection of 1,065 nonredundant giant virus MAGs was established to represent giant virus populations in Uranouchi Inlet. These genomes were then used as the species-level references for microdiversity analyses.

The PIMA approach was performed to assess the quality of 1,065 representative giant virus MAGs (Fig. S4a; Tables S2, S3). A total of 68 MAGs were not evaluated by PIMA because they were classified as long branches and did not belong to any clade within our designated threshold for assessment. The quality of these MAGs was estimated using checkV (Table S3). Among the other 997 MAGs, the median consistency and redundancy were 83.33% and 14.81%, respectively (Fig. S4a). A notable proportion of MAGs (N = 392) exhibited consistency greater than 80% and redundancy less than 20% (see Supplementary text), which indicated strong performance of the pipeline when applied to the Uranouchi metagenomes.

Community Composition of Giant Virus MAGs

We conducted taxonomic classification by reconstructing a phylogenomic tree using concatenated amino acid sequences of three marker genes: RNA polymerase alpha subunit (RNAPa), RNA polymerase beta subunit (RNAPb), and DNA polymerase B (DNApolB) (Fig. 1a). Out of 1,065 giant virus MAGs, 1,052 were taxonomically categorized within known orders of *Nucleocytoviricota*, including *Imitervirales*, *Algavirales*, *Asfuvirales*, *Pimascovirales*, and *Pandoravirales*. The proportions of the number of viruses in different lineages observed in Uranouchi Inlet showed a similar but slightly different trend compared with those noted in global oceanic surveys. Of the 1,052 *Nucleocytoviricota* MAGs, 831 were classified into the order *Imitervirales* (78.99%), which is a higher proportion than was detected in the *Tara* Oceans project (66.19%)[51] and indicates a greater diversity of hosts for *Imitervirales* in the coastal waters. The size of the giant virus MAGs ranged from 200 Kbp to 1.9 Mbp. We discovered three clades represented by 22 MAGs that together formed a sister clade close to the *Ectocarpus siliculosus* virus group, which includes viruses that infect brown algae (Fig. S5). One of the clades (CladeC in Fig. S5; N = 12) was characterized by large genomes ranging from 1.2 Mbp to 1.9 Mbp (median = 1.7 Mbp).

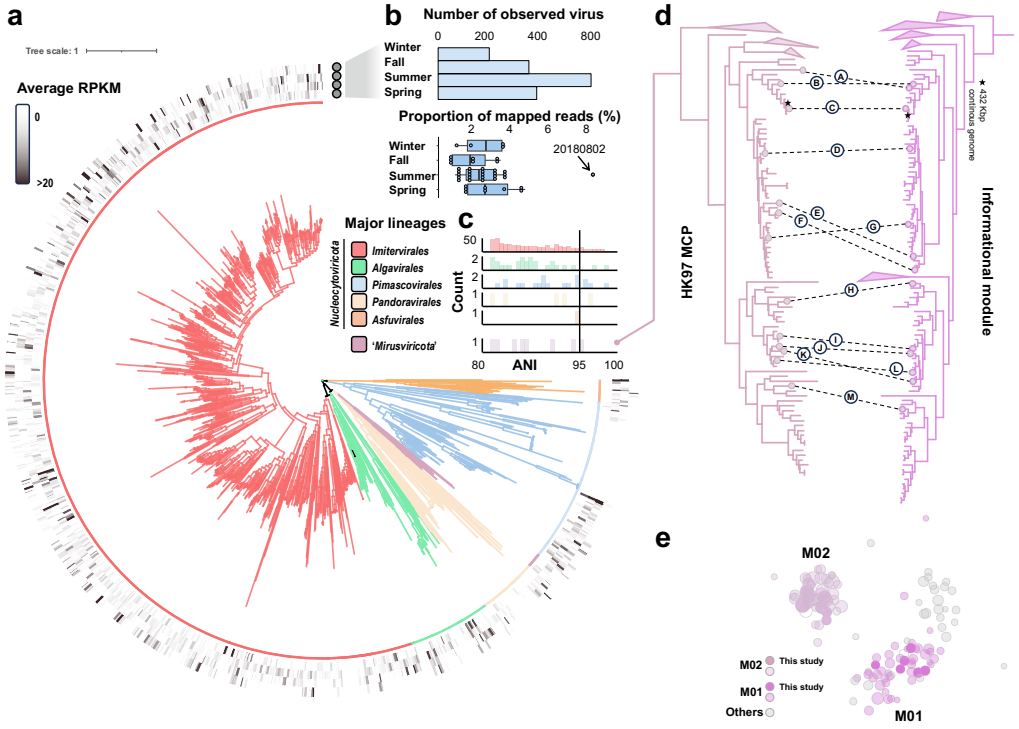

Figure 1. Community composition of giant viruses in Uranouchi Inlet. (**A**) Phylogenetic tree of 1,065 MAGs. The maximum likelihood tree, constructed from a concatenated alignment of three hallmark genes (DNApolB, RNApolA, and RNApolB), showcases the diversity among giant viruses. The tree is color coded to represent six lineages used in subsequent analyses: *Imitervirales*, *Algavirales*, *Pandoravirales*, *'Mirusviricota,' Pimascovirales*, and *Asfuvirales.* (**B**) Number of observed giant virus MAGs and the proportion of mapped reads to giant viruses relative to all sequenced reads among four seasons. (**C**) Number of MAGs with high ANI to the genomes in three public global genome datasets[27–29]. (**D**) Maximum likelihood phylogenetic tree built from the '*Mirusviricota'* MAGs based on a concatenated alignment of three hallmark informational genes (RNApolA, RNApolB, DNApolB), right; maximum likelihood phylogenetic tree built from the '*Mirusviricota*' MAGs of the HK97-fold major capsid protein, left. Dashed lines on the two phylogenetic trees indicate the positions of 13 MAGs identified in Uranouchi Inlet. (**E**) Bipartite networks of '*Mirusviricota'* MAGs. The two colored groups represent two families of *Mirusviricota*, M01 and M02, and gray circles represent five other families defined in a previous study[29]. The darker circles represent the MAGs assembled in this study, whereas the lighter circles are from the previous study.

Additionally, we identified 13 MAGs affiliated with the recently identified viruses in the phylum '*Mirusviricota*' (Fig. 1a)[29]. The 13 MAGs encoded a nearly complete set of core genes commonly found in previously discovered mirusviruses. The largest mirusvirus MAG from Uranouchi Inlet had a genome size exceeding 484 Kbp (UUJ171113_122), which surpassed all previously discovered mirusvirus genomes. One MAG, UUJ180313_81, showed 86% ANI and wide alignment breadth with a nearly complete mirusvirus genome detected in the Mediterranean Sea (Fig. S4b). The size of this MAG (>398 Kbp) was almost as large as the nearly complete continuous mirusvirus genome (432 Kbp)[29], which demonstrated that the assembly and binning process was appropriate for detecting mirusvirus genomes. Reconstructing the phylogenomic tree of these 13 MAGs with previous mirusviruses revealed that six MAGs from Uranouchi Inlet corresponded to family M01, whereas the other seven MAGs were classified in family M02 (Fig. 1d). M01 and M02 are the most diverse of the seven recognized families of the '*Mirusviricota*' phylum[29]. This classification was also supported by the gene content of those MAGs (Fig. 1e). Additionally, the presence of genes encoding the HK97-fold major capsid protein in the Uranouchi MAGs supported their identification as mirusviruses. In the following analyses, we used five *Nucleocytoviricota* orders, *Imitervirales*, *Algavirales*, *Asfuvirales*, *Pimascovirales*, and *Pandoravirales*, and one viral phylum, '*Mirusviricota*,' as the main lineages.

Community and Population Dynamics of Giant Viruses

Giant virus community diversity (Fig. 2a) and composition at the main lineage level (Fig. 2b) showed variations during the sampling period from January 2017 to September 2018. Lineage-level diversity increased during the summer months across two years (Fig. 2b), which indicated the presence of a consistent annual pattern within the giant virus community. Overall, *Imitervirales* were constantly abundant at any given time point and constituted the majority of the giant virus community all the time, followed by *Pimascovirales* and *Algavirales*. However, the abundance of many giant virus lineages demonstrated distinct seasonal dynamics (Fig. S6a). Specifically, *Imitervirales* displayed a relatively persistent presence throughout the two-year period, whereas enrichment of other giant virus lineages occurred during the summer months (i.e., June, July, and August); this indicated seasonal flourishing of a diverse range of their hosts. Among the top 50 most abundant giant virus MAGs (cumulative RPKM of all samples for each MAG), the majority belonged to the orders *Imitervirales*, *Pimascovirales*, and *Algavirales* (Fig. S7a). However, the most abundant MAG, *Heterocapsa circularisquama* DNA virus (HcDNAV)[52], which belongs to *Asfuvirales* and infects the toxic bloom-forming dinoflagellate *Heterocapsa circularisquama*, showed remarkable abundance in July 2018 (maximum RPKM, 992.71), which accounted for 35.2% of the total viral community.

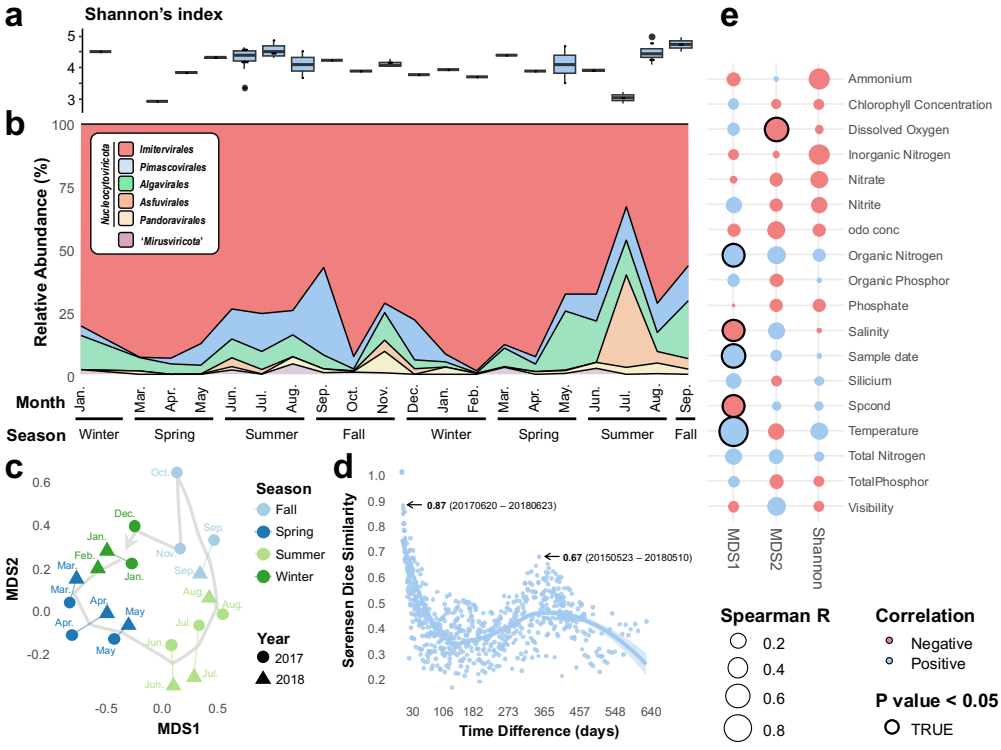

Figure 2. Time-series variation of giant virus community in Uranouchi Inlet. (**A**) Shannon's diversity index for the total giant virus community. (**B**) The composition of the viral community at the lineage level. (**C**) Non-metric

multidimensional scaling (NMDS) ordination plot based on Sorensen–Dice dissimilarity between communities from January 2017 to September 2018. The stress value of the NMDS was 0.03. (**D**) Pairwise community similarity, which compares the similarity between viral communities from two samples, analyzed over time intervals ranging from 0 to 640 days. The community similarity level is given by Sorensen–Dice dissimilarity based on the presence/absence of species in the community. (**E**) Spearman's correlation coefficients between environmental factors and viral community. Correlation analyses were performed for MDS1 and MDS2 of panel (**C**) and Shannon's index of the viral community. P-values were adjusted by the Benjamini–Hochberg procedure.

NMDS ordination of the datasets revealed a clear year-round cycle of the entire giant virus community (Fig. 2c), which demonstrated month-to-month succession from January to December. Year-round recovery of the community was also demonstrated by a community similarity analysis (Fig. 2d). The pairwise community similarity showed a cyclical yearly pattern with a peak at an approximately 365-day interval. Two samples from May across the two years exhibited a similarity index of 0.67. Temperature and dissolved oxygen were the two most significant environmental variables related to the year-round recovery of the viral community (Fig. 2e).

At the population level, the occurrence frequency of giant virus MAGs over a period of 20 months exhibited a right-skewed distribution pattern, with 90.06% of MAGs appearing in fewer than five months. MAGs that were present for five or more months were predominantly from *Imitervirales*, *Pimascovirales*, and *Algavirales* (Fig. S7b). Only 21 MAGs were present in four seasons (Fig. 3a). Distinct distribution patterns were evident in the dynamics of giant virus MAGs. Based on the population dynamic patterns, giant virus MAGs that had appeared before the second winter (December 2017) (N = 744) were categorized into three niche groups (persistent [n = 9], seasonal [n = 389], sporadic [n = 318) or as 'other' (n= 28) (Fig. 3; see Methods). The proportion of MAGs that belonged to these categories are shown in Fig. 3b. Among all the six lineages, *Algavirales* was the only seasonal population that showed a dominant proportion (68.9%). *Imitervirales* and *Pimascovirales* included persistent viral populations. *Asfuvirales* had the highest proportion of sporadic populations (75%). The typical dynamic pattern of each category is provided in Fig. 3c.

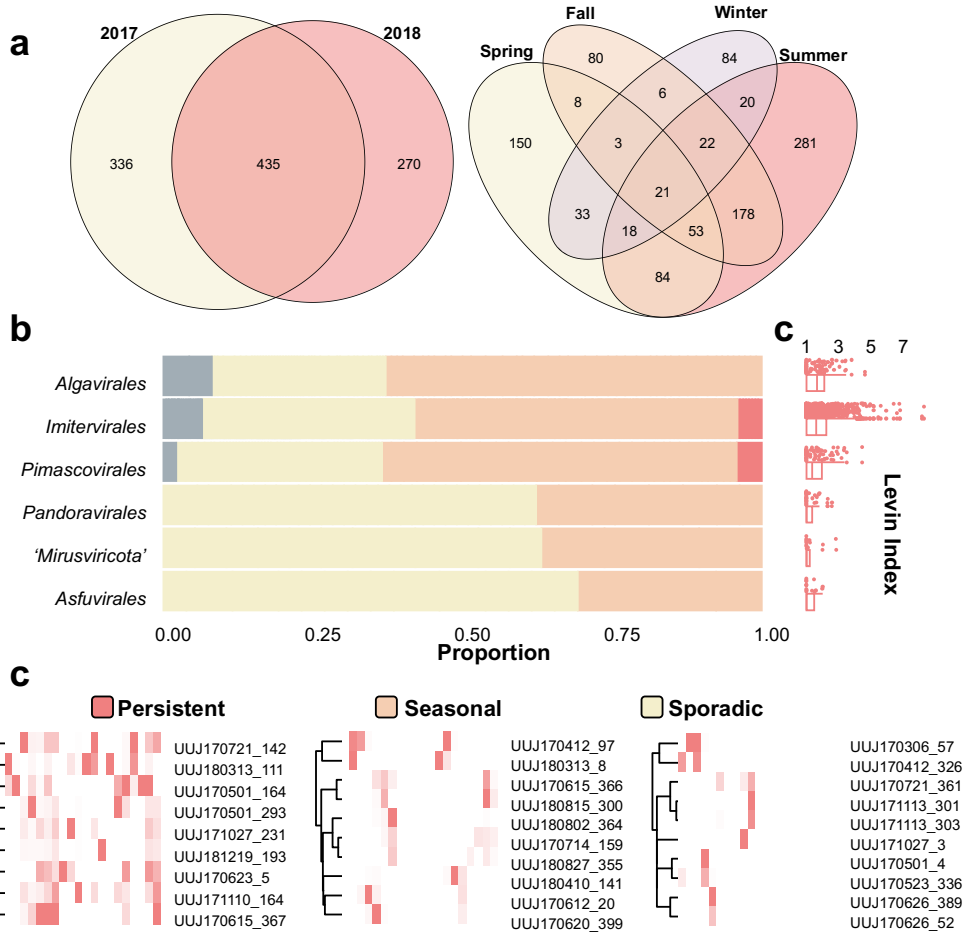

Figure 3. Three niche categories of giant virus populations. (**A**) Venn diagrams showing the breakdown of MAG occurrences in different periods. (**B**) Distribution of MAGs in the niche categories for the six major lineages. (**C**) Abundance heatmap for the persistent populations (left), seasonal populations (middle), and sporadic populations (right). The x-axis represents 20 months from 2017 to 2018 and the y-axis represents the randomly chosen MAG examples.

Persistence and Microdiversity

Then, we investigated factors that are associated with niche breadth (persistence) measured by the Levins' index (Fig. 4a). The strongest correlation was observed between the Levins' index and the average ND (Spearman's correlation coefficient 0.41, P-value < 0.001) (Fig. 4A). Consistent with this observation, the average ND values of persistent and seasonal giant virus populations were higher than that of sporadic populations (Fig. S8a). Similarly, the average ND of MAGs appearing in multiple seasons tended to be higher than that of MAGs occurring in only one season (Fig. S8Bb). Because the viral community exhibited a year-round recovery pattern, we investigated whether the

recurrence of individual MAGs was related to their microdiversity. The average ND of recurrent MAGs (those present in both 2017 and 2018) was significantly higher compared with that of non-recurrent MAGs (P-value < 0.001) (Fig. S8c). Overall, giant viruses with higher persistence levels displayed higher ND (Fig. 4c). At the lineage level, *Algavirales* and *Imitervirales* had the highest median ND, followed by '*Mirusviricota*,' *Pimascovirales*, *Pandoravirales*, and *Asfuvirales* (Fig. 4b). All lineages exhibited higher ND in the niche categories associated with generalists (persistent and seasonal) compared with that of specialists (sporadic). However, '*Mirusviricota*' and *Pandoravirales* showed no significant correlation between their ND and niche breadth (Fig. S9d).

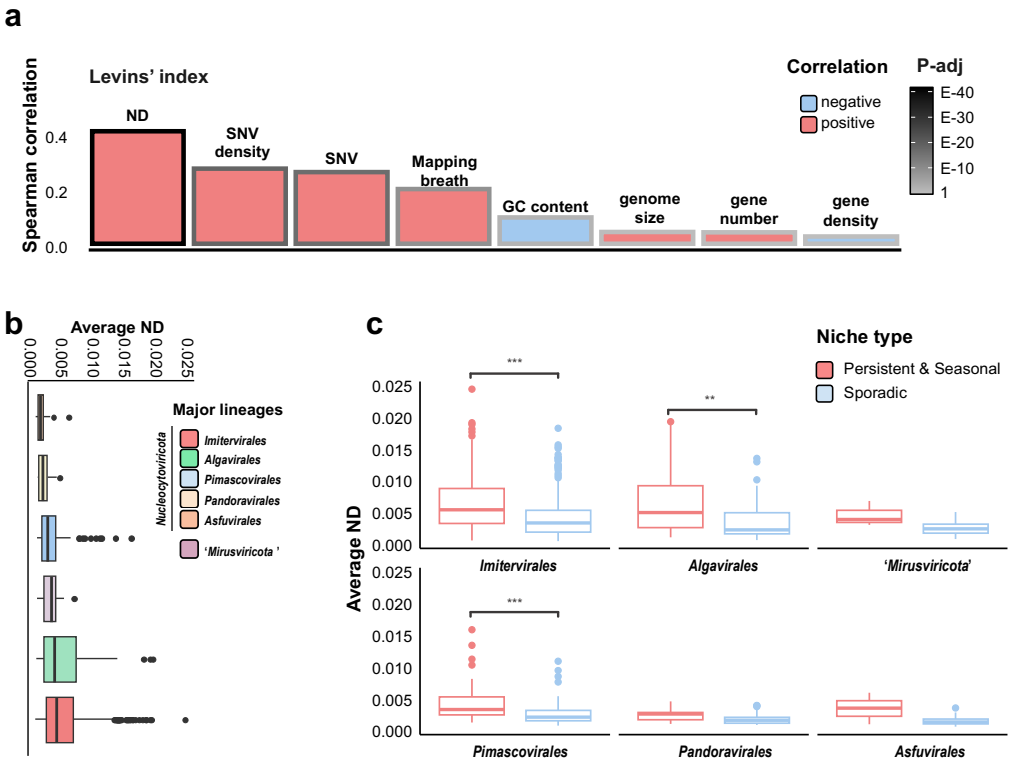

Figure 4. Persistence and microdiversity. (**A**) Correlation between niche breadth (Levins' index) and features of giant virus genomes. ND, SNV density, and SNV count were obtained by InStrain v1.0.0. The frame color of the bar plots represents the adjusted P-value (Benjamini–Hochberg procedure). (**B**) Average ND of each lineage. Lineages are sorted by median values. (**C**) Comparison of average ND between generalists (persistent and seasonal) and specialists (sporadic) of each viral lineage. The Wilcoxon rank-sum test was used to determine the significance of comparison. For box plots, center lines show the medians, box limits represent the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, and outliers are represented by dots.

The microdiversity of giant viruses, assessed through ND and SNV/Mb, was not strongly correlated with genome size (R^2 = 0.02 and 0.06, respectively; Fig. 4a, S9c). Additionally, to address

potential biases in microdiversity detection due to sequencing depth and mapping approaches, we analyzed the relationship between microdiversity and the read coverage of the analyzed sample (i.e., the highest RPKM across samples). ND and SNV/Mb were not influenced by coverage depth (R^2 = 0.02 and 0.05, respectively) as much as niche breadth (Fig. S9a, b). Both results indicate that our microdiversity measurement was not significantly influenced by possible artifacts in binning and mapping.

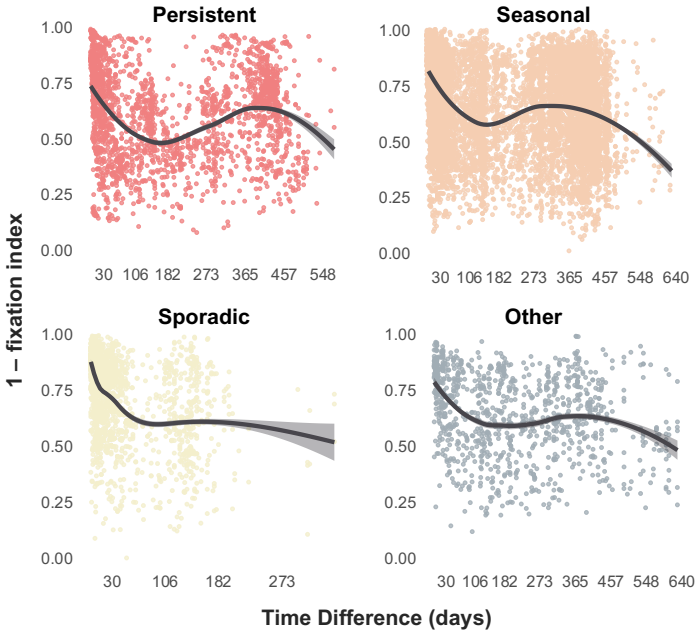

Figure 5. Pairwise microdiversity similarity calculated over a 20-month interval that compares the microdiversity of each MAG from every two samples. The pairwise microdiversity similarity level was estimated by (1 − fixation index). The plot is drawn for the viral MAGs of four niche categories: persistent, seasonal, sporadic, and other, respectively. Locally estimated scatterplot smoothing (LOESS) is used to demonstrate the tendency.

Finally, we explored the temporal dynamics of microdiversity within viral populations. First, we calculated the fixation index of each viral population (i.e., MAG) across 42 samples in 20 months to assess the pairwise distance of the genetic structure of the viral population among samples. Similar to the community level analysis, overall year-round recovery patterns of microdiversity were observed in both the persistent and seasonal categories (Fig. 5). The pairwise microdiversity similarity of both categories exhibited a yearly cyclical pattern with a peak at approximately 365-day intervals. However, for the viral populations in the other two ecological groups, a tendency of no or weak recovery was observed. Subsequently, we identified two example patterns of microdiversity dynamics for individual populations in the persistent group (Fig. 6a, b;

Fig. S10). In certain populations, temporal shifts in allele frequencies were observed, with many SNV sites being dominated by a single allele at specific time points (e.g., UUJ170721_142) (Fig. S10). UUJ170721_142 also demonstrated a clear year-round recovery (Fig. 6b), with seasonal ecotypes (subpopulations) that shared similar microdiversity within the same seasons (Fig. 6c). Despite the year-round recovery, the ND values for this viral population in each month showed no relationship with the ecotype clusters (Fig. 6b). Moreover, certain viral populations exhibited relatively stable allele frequencies throughout the months of their occurrence across the two years of analysis (e.g., UUJ170623_5, *Pimascovirales*) (Fig. 6a, b). Additionally, some giant viruses displayed a high degree of variation, and most of the populations from different samples had alleles that differed from the ones in the reference MAG. One example of this pattern was observed in UUJ180313_111, which is phylogenetically close to Organic Lake *Phycodnavirus* (Fig. S10).

Discussion

Metagenomics has largely improved our understanding of giant viruses by revealing their distribution across various biomes worldwide using datasets assembled from global samples[27–29,53]. The usage of genomes from metagenomic assemblies is widely accepted, but different approaches can lead to different interpretations[54]. This is particularly crucial for virus studies because viruses exhibit high diversity and there are only a few reference genomes available compared with environmental data[55]. A viral MAG is a consensus that masks microdiversity and may not represent any specific genotype in the environment. To deeply analyze viral genomes, such as diversity at the intra-species level, it is critical that the representative genome of each population is high quality. One of the best solutions is manually curating the viral bins using interactive metagenomic tools, like Anvi'o[29,56]. However, this approach demands considerable time, labor, and expertise. To improve research efficiency and reproducibility, there is a need for a pipeline that includes automated curation and refinement processes to substitute manual checking. To address this, we developed a dedicated metagenomic pipeline for recovering MAGs of giant viruses (Fig. S2), which specifically focuses on removing potential contamination from cellular organisms and eliminating chimeric bins (Fig. S2). Subsequently, to overcome the shortage of reference genomes, we developed a phylogeny-informed quantitative assessment approach based on the principle that evolutionarily related viruses tend to have similar gene contents. Validation using this novel method demonstrated that most of the MAGs generated in this study were high quality, which indicated that the pipeline was efficient for quality control. The MAGs in this study covered all known main lineages (Fig. 1) other than

Chitovirales, which are not widely distributed and abundant in marine environments[57]. Moreover, the automatic pipeline generated nearly complete mirusvirus genomes[29], which demonstrated that the giant virus screening threshold had high sensitivity for detecting novel giant virus lineages. Overall, our genome dataset has high reliability for microdiversity studies. Furthermore, the pairwise ANI pattern of giant viruses (Fig. S3) was similar to that observed in bacteria[50], which supported the concept of 'species' for giant viruses, with 95% ANI as an approximate species boundary as suggested by previous studies[33,34].

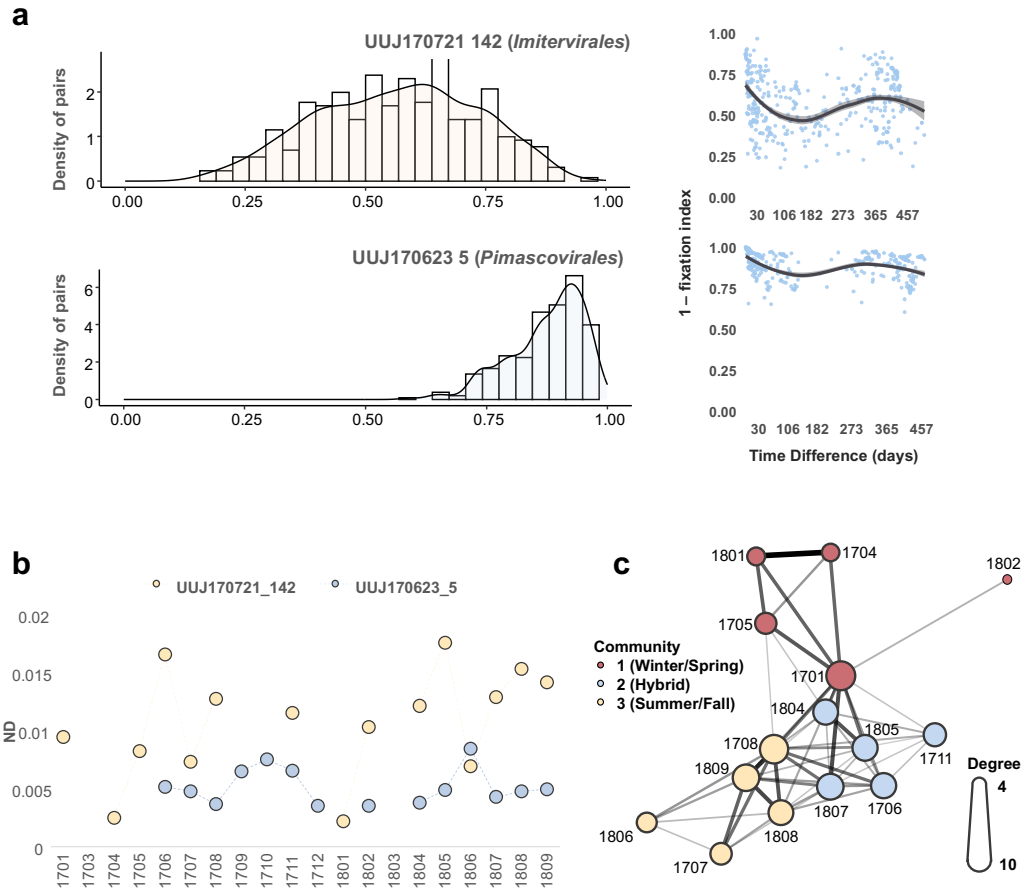

Figure 6. Temporal dynamics of SNV allele frequencies within different populations. (**A**) Histogram of similarity values (1 − fixation index) for two MAG samples (left panel). Pairwise microdiversity similarity comparing the microdiversity of each MAG between every two samples of three MAGs (right panel). The pairwise microdiversity similarity level was estimated by 1 − fixation index. LOESS was used to show the trend of year-round recovery. (**B**) ND dynamics of the two MAGs across 20 months. (**C**) The network of the pair-wised monthly microdiversity similarity of UUJ170721_142 was based on similarity value (1 − fixation index), which is represented by edge width. The size of the nodes represents the degree of the node, and the color represents identified communities. Community search was performed using the R package "igraph" with the function "cluster_louvain." The structure of the network was determined by the "igraph"

package with the function "layout_with_fr" and visualized using the "ggraph" package. Only similarities greater than 0.5 were used for the network analysis.

Uranouchi Inlet, Japan is a semi-enclosed eutrophic inlet with high biodiversity, from unicellular organisms to large animals (Fig. S1)[58,59]. A previous study[16] identified year-round recovery of the *Imitervirales* community based on amplicon sequence variants of a single marker gene, DNApolB, in Uranouchi Inlet. Another recent study revealed clear seasonal dynamics of giant virus communities in the photic and aphotic layers of a freshwater lake[60]. Although a few additional studies have explored the temporal dynamics of giant viruses[61–63], no study has addressed the seasonality of the whole giant virus community at the population level. In this study, we characterized the temporal dynamics of the giant virus community across all environmental giant viral lineages. Consistent year-round recovery was observed throughout two years (Fig. 2c, d). In addition to five orders of the phylum *Nucleocytoviricota*, we also revealed seasonality of two families (M1 and M2) that belong to a recently discovered phylum, '*Mirusviricota*' (Fig. S6b); this supports a ubiquitous distribution of mirusviruses in marine environments[29]. Overall, only a few MAGs (N = 21) were present across all seasons (Fig. 3a), which indicates that most of the giant viruses have seasonal preferences. To clearly demonstrate the niche breadth, or ecological strategy, we categorized the viral MAGs into one of three categories: persistent, seasonal, or sporadic (Fig. 3b). The proportion of niche categories varied across lineages, which likely resulted from their distinctive host ranges, as viruses can only thrive when their hosts thrive. For example, a large proportion of seasonal populations was observed in *Algavirales* (Fig. 3b), which may be because they mainly infect algal species exhibiting seasonality[64]. On the contrary, *Pimascovirales* also showed abundance peaks across two years (Fig. S5), but they displayed more sporadic occurrence than *Algavirales* (Fig. 3b). This may be primarily because pimascoviruses infect large animals, such as fishes[65], and thus are likely carried to Uranouchi Inlet by these swimming organisms. Overall, similar to the heterogeneity observed in their spatial distributions[11,51], different lineages of giant viruses also exhibited temporal variations in their distributions. The presence of persistent and seasonal giant virus populations primarily contributed to the year-round recovery of the viral community (Fig. 2; Fig. 3). More importantly, we observed that the viral populations of persistent and seasonal categories also exhibited year-round recovery in their intra-population genetic structure. Therefore, seasonal changes were seen at both viral community and population levels. This demonstrates that the

dynamics at the population level, such as seasonal ecotypes (Fig. 6D), may be an important factor that contributes to shaping the seasonal dynamics of the viral community.

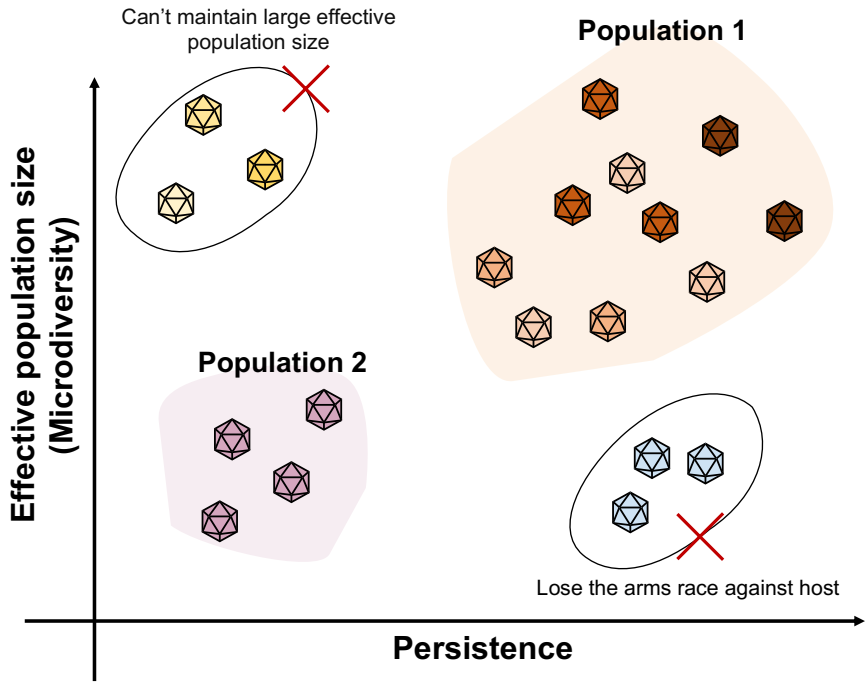

Figure 7. Summary of the relationships among microdiversity, effective population size, and persistence of giant viruses.

Our study revealed a trend that sporadic and persistent GVs showed comparatively low and high microdiversity, respectively (Fig. 4; Fig. S8). Theoretically, the level of microdiversity positively correlates with the effective population size (the number of individuals that effectively participate in producing the next generation) under the assumption that mutation rates are comparable across populations and most of the variations are neutral[66,67]. Reduced effective population size amplifies the impact of genetic drift and leads to low diversity, whereas elevated effective population size allows for accumulation of a higher level of neutral mutations within the population and enhances the efficiency of natural selection[68]. Notably, similar trends between the level of genetic variation and ecological dynamics have been reported for prokaryotes and eukaryotes in aquatic environments[6,69–71]. Based on the above theoretical framework, those studies hypothesized that increased genetic diversity in cellular organisms was associated with adaptability to specific micro-niches, defense against viruses[69], stability in abundance[70,71], and lack of recent population bottleneck[70,71].

Because viruses are obligate parasites of their hosts, specific factors could account for the relationship between the microdiversity and persistence of viruses, and explain why sporadic viruses rarely show high microdiversity and persistent viruses rarely show low microdiversity (Fig. 7). Viral

particles lose infectivity over time. For example, in relatively severe light conditions, the average loss rate of viral infectivity was 0.2 h^{-1}[72]. Therefore, sporadic viruses may undergo the bank model of virus–host interactions[73] and experience a genetic bottleneck that leads to microdiversity loss (Fig. 7). In contrast, for viruses to void decay and to persist in an environment, they need to recurrently infect hosts that also persist[74,75]. Based on previous observations[6,69–71], these persistent hosts tend to possess a large effective population size, which results in a higher rate of fixation of advantageous mutations, including defense mechanisms against viruses[76]. In this situation, viruses with large population sizes are advantageous because they can rapidly acquire advantageous traits to survive under the severe virus–host arms race. Persistent viruses with a small effective population size may become losers in this co-evolutionary arms race.

However, testing of our hypothesis is limited because we only analyzed virus metagenomics during a 20-month observational period. Moreover, in addition to virus–host interactions, competitive exclusion during co-infection might also influence the differences in microdiversity levels within giant virus populations. Despite these limitations, this research represents the first comprehensive analysis of temporal dynamics and microdiversity in giant viruses. The observed lineage-specific microdiversity provides a novel perspective on the varied ecological and evolutionary processes that affect viral lineages. This study both highlights the critical role of virus–host interactions in shaping the dynamics of giant virus populations and establishes an essential framework for understanding the intricate relationships between these distinct viral entities and their host communities. Populations of giant viruses with high microdiversity may enhance host resource specialization, potentially benefiting nutrient recycling and promoting long-term system stability.

Author Contributions

YF and LM performed most of the bioinformatics analyses in this study. JX and YO contributed to the bioinformatics analyses. KN performed sampling and LM performed DNA extraction. YG and TH contributed to sequencing. HE and HO designed the study. HE, YO, and HO co-supervised YF. YF generated the initial draft and LM improved it. All authors contributed to the interpretation of data and writing of the manuscript, and all approved the final draft.

Acknowledgments

This work was supported by JSPS/KAKENHI (Nos. 21H05057, 22H00384, 22H00385, 16H06279 [PAGS]), Scientific Research on Innovative Areas from the Ministry of Education, Culture, Science, Sports and Technology (MEXT) of Japan (Nos. 16H06429, 16K21723, 16H06437), The Kyoto University Foundation, and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (Nos. 2021-33, 2019-33, 2018-31, 2017-25). Computational work was completed at the SuperComputer System, Institute for Chemical Research, Kyoto University. We thank Mallory Eckstut, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature* **393**, 464–467 (1998).
2. Fuhrman, J. A. & Campbell, L. Microbial microdiversity. *Nature* **393**, 410–411 (1998).
3. Schloter, M., Lebuhn, M., Heulin, T. & Hartmann, A. Ecology and evolution of bacterial microdiversity. *FEMS Microbiol Rev* **24**, 647–660 (2000).
4. Riehle, M. M., Bennett, A. F. & Long, A. D. Genetic architecture of thermal adaptation in Escherichia coli. *Proceedings of the National Academy of Sciences* **98**, 525–530 (2001).
5. Thompson, J. R. *et al.* Genotypic diversity within a natural coastal bacterioplankton population. *Science (1979)* **307**, 1311–1313 (2005).
6. García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C. & Puente-Sánchez, F. Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *ISME J* **13**, 2969–2983 (2019).
7. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* **11**, 6013 (2020).
8. Boon, M., Holtappels, D., Lood, C., van Noort, V. & Lavigne, R. Host range expansion of Pseudomonas virus LUZ7 is driven by a conserved tail fiber mutation. *Phage* **1**, 87–90 (2020).
9. Suttle, C. A. Marine viruses - Major players in the global ecosystem. *Nat Rev Microbiol* **5**, 801–812 (2007).
10. Weitz, J. S. & Wilhelm, S. W. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol Rep* **4**, (2012).
11. Endo, H. *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* **4**, 1639–1649 (2020).
12. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME Journal* **7**, 1678–1695 (2013).
13. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat Commun* **9**, 373 (2018).

14. Kaneko, H. *et al.* Eukaryotic virus composition can predict the efficiency of carbon export in the global ocean. *iScience* **24**, 102002 (2021).
15. Martínez, J. M., Schroeder, D. C., Larsen, A., Bratbak, G. & Wilson, W. H. Molecular dynamics of Emiliania huxleyi and cooccurring viruses during two separate mesocosm studies. *Appl Environ Microbiol* **73**, 554–562 (2007).
16. Prodinger, F. *et al.* Year-round dynamics of amplicon sequence variant communities differ among eukaryotes, Imitervirales and prokaryotes in a coastal ecosystem. *FEMS Microbiol Ecol* **97**, fiab167 (2021).
17. Sullivan, M. B. Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. *J Virol* **89**, 2459–2461 (2015).
18. Achtman, M. & Wagner, M. Microbial diversity and the genetic nature of microbial species. *Nat Rev Microbiol* **6**, 431–440 (2008).
19. Endo, H., Ogata, H. & Suzuki, K. Contrasting biogeography and diversity patterns between diatoms and haptophytes in the central Pacific Ocean. *Sci Rep* **8**, 10916 (2018).
20. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Preprint at (2010).
21. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
22. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
23. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
24. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
25. Li, H. *et al.* The sequence alignment/map format and SAMtools. *bioinformatics* **25**, 2078–2079 (2009).
26. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
27. Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
28. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun* **11**, 1–11 (2020).
29. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature* 1–7 (2023).
30. Subramaniam, K. *et al.* A new family of DNA viruses causing disease in crustaceans from diverse aquatic biomes. *mBio* **11**, 1–14 (2020).
31. Yutin, N., Wolf, Y. I., Raoult, D. & Koonin, E. V. Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J* **6**, 1–13 (2009).
32. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864–2868 (2017).
33. Zhang, R., Takemura, M., Murata, K. & Ogata, H. "Mamonoviridae", a proposed new family of the phylum Nucleocytoviricota. *Arch Virol* **168**, 80 (2023).
34. Aylward, F. O. *et al.* Taxonomic update for giant viruses in the order Imitervirales (phylum Nucleocytoviricota). *Arch Virol* **168**, 283 (2023).

35. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).
36. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
37. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534 (2020).
38. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol* **19**, 1–18 (2021).
39. Koonin, E. V. & Yutin, N. *Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. Advances in Virus Research* vol. 103 (Elsevier Inc., 2019).
40. Woodcroft, B. J. & Newell, R. CoverM: Read coverage calculator for metagenomics. Preprint at (2017).
41. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
42. Oksanen, J. *et al.* vegan: Community Ecology Package. R package. version 2.5-3. *https://CRAN.R-project.org/package=vegan* (2018).
43. Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat Biotechnol* **39**, 727–736 (2021).
44. Nei, M. & Li, W.-H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences* **76**, 5269–5273 (1979).
45. Sjöqvist, C., Delgado, L. F., Alneberg, J. & Andersson, A. F. Ecologically coherent population structure of uncultivated bacterioplankton. *ISME J* **15**, 3034–3049 (2021).
46. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York* **ISBN 978-3**, (2016).
47. Nishimura, Y., Yamada, K., Okazaki, Y. & Ogata, H. DiGAlign: Versatile and Interactive Visualization of Sequence Alignment for Comparative Genomics. *Microbes Environ* **39**, ME23061 (2024).
48. Paul Shannon *et al.* Cytoscape: A Software Environment for Integrated Models. *Genome Res* **13**, 426 (2003).
49. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–W296 (2021).
50. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).
51. Meng, L. *et al.* Genomic adaptation of giant viruses in polar oceans. *Nat Commun* **14**, 6233 (2023).
52. Ogata, H. *et al.* Remarkable sequence similarity between the dinoflagellate-infecting marine girus and the terrestrial pathogen African swine fever virus. *Virol J* **6**, 1–8 (2009).
53. Schulz, F. *et al.* Hidden diversity of soil giant viruses. *Nat Commun* **9**, 1–9 (2018).
54. Delmonta, T. O. & Eren, A. M. Simulations predict microbial responses in the environment? This environment disagrees retrospectively. *Proc Natl Acad Sci U S A* **114**, E8947–E8949 (2017).
55. Roux, S. *et al.* Minimum information about an uncultivated virus genome (MIUVIG). *Nat Biotechnol* **37**, 29–37 (2019).
56. Eren, A. M. *et al.* Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ* **2015**, 1–29 (2015).
57. Gjessing, M. C. *et al.* Salmon gill poxvirus, the deepest representative of the Chordopoxvirinae. *J Virol* **89**, 9348–9367 (2015).

58. Jaysankar, D., Fukami, K., Iwasaki, K. & Okamura, K. Occurrence of heavy metals in the sediments of Uranouchi Inlet, Kochi prefecture, Japan. *Fisheries Science* **75**, 413–423 (2009).
59. Tarutani, K., Nagasaki, K. & Yamaguchi, M. Viral impacts on total abundance and clonal composition of the harmful bloom-forming phytoplankton: Heterosigma akashiwo. *Appl Environ Microbiol* **66**, 4916–4920 (2000).
60. Zhang, L., Meng, L., Fang, Y., Ogata, H. & Okazaki, Y. Spatiotemporal dynamics revealed the dark water community of giant virus from a deep freshwater lake. *bioRxiv* 2004–2024 (2024).
61. Gran-Stadniczeñko, S. *et al.* Seasonal dynamics of algae-infecting viruses and their inferred interactions with protists. *Viruses* **11**, 1043 (2019).
62. Roux, S. *et al.* Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat Commun* **8**, 858 (2017).
63. Tomaru, Y., Tarutani, K., Yamaguchi, M. & Nagasaki, K. Quantitative and qualitative impacts of viral infection on a Heterosigma akashiwo (Raphidophyceae) bloom in Hiroshima Bay, Japan. *Aquatic Microbial Ecology* **34**, 227–238 (2004).
64. Alegria Zufia, J., Farnelid, H. & Legrand, C. Seasonality of coastal picophytoplankton growth, nutrient limitation, and biomass contribution. *Front Microbiol* **12**, 786590 (2021).
65. Sun, T.-W. W. *et al.* Host Range and Coding Potential of Eukaryotic Giant Viruses. *Viruses* **12**, 1337 (2020).
66. Chen, Z. *et al.* Prochlorococcus have low global mutation rate and small effective population size. *Nat Ecol Evol* **6**, 183–194 (2022).
67. Wang, X. & Feng, X. Challenges in estimating effective population sizes from metagenome-assembled genomes. *Front Microbiol* **14**, 1331583 (2024).
68. Batut, B., Knibbe, C., Marais, G. & Daubin, V. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat Rev Microbiol* **12**, 841–850 (2014).
69. Needham, D. M., Sachdeva, R. & Fuhrman, J. A. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J* **11**, 1614–1629 (2017).
70. Okazaki, Y., Nakano, S., Toyoda, A. & Tamaki, H. Long-read-resolved, ecosystem-wide exploration of nucleotide and structural microdiversity of lake bacterioplankton genomes. *mSystems* **7**, e00433-22 (2022).
71. Meziti, A. *et al.* Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. *ISME J* **13**, 767–779 (2019).
72. Mojica, K. D. A. & Brussaard, C. P. D. Factors affecting virus dynamics and microbial host–virus interactions in marine environments. *FEMS Microbiol Ecol* **89**, 495–515 (2014).
73. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* **13**, 278–284 (2005).
74. Ignacio-Espinoza, J. C., Ahlgren, N. A. & Fuhrman, J. A. Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat Microbiol* **5**, 265–271 (2020).
75. Castledine, M. & Buckling, A. Critically evaluating the relative importance of phage in shaping microbial community composition. *Trends Microbiol* (2024).
76. Gossmann, T. I., Keightley, P. D. & Eyre-Walker, A. The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol* **4**, 658–667 (2012).