Dataset of Oyster Virome and the Remarkable Virus Diversity in Filter-Feeding Oysters

Jing-Zhe Jiang (jingzhejiang@gmail.com)
South China Sea Fisheries Research Institute https://orcid.org/0000-0001-5260-7822

Yi-Fei Fang
South China Sea Fisheries Research Institute

Hong-Ying Wei
South China Sea Fisheries Research Institute

Ying-Xiang Guo
South China Sea Fisheries Research Institute

Li-Ling Yang
South China Sea Fisheries Research Institute

Tao Jin
Guangdong Magigene Biotechnology Co., Ltd

Mang Shi
Sun Yat-Sen University

Shao-Kun Shi
Shenzhen Fisheries Development Research Center

Meng Wang
bureau of agriculture and rural affairs of conghua district

Tuo Yao
South China Sea Fisheries Research Institute

Jie Lu
South China Sea Fisheries Research Institute

Ling-Tong Ye
South China Sea Fisheries Research Institute

Ming Duan
Chinese academy of sciences

Dian-Chang Zhang
south china sea fisheries research institute

Research

Keywords: Crassostrea hongkongensis, Bivalve, Mollusk, Metagenome, Circoviridae, Viral-Like 51 Particle Enrichment, South China, Multiple Displacement Amplification, phi29

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1044974/v1

Dataset of oyster virome and the remarkable virus diversity in filter-feeding oysters

Jing-Zhe Jiang*[#, a], Yi-Fei Fang[#, a, h], Hong-Ying Wei[a, c], Ying-Xiang Guo[a, e], Li-Ling Yang[a, e], Tao Jin[c], Mang Shi[d], Shao-Kun Shi[f], Meng Wang[g], Tuo-Yao[a], Jie Lu[a], Ling-Tong Ye[a], Ming Duan*[, b], Dian-Chang Zhang*[, a]

a. Key Laboratory of South China Sea Fishery Resources Exploitation and Utilization, Ministry of Agriculture, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510300, Guangdong, China
b. State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, Hubei, China
c. Guangdong Magigene Biotechnology Co., Ltd, Guangzhou 510000, Guangdong, China
d. School of Medicine, Sun Yat-sen University, Shenzhen 518107, Guangdong, China
e. Tianjin Agricultural University, Tianjin 300384, China
f. Shenzhen Fisheries Development Research Center, Shenzhen 518067, Guangdong, China
g. Bureau of Agriculture and Rural Affairs of Conghua District, Guangzhou 510925, Guangdong, China
h. Current address: Shanghai Majorbio Bio-Pharm Technology Co., Ltd, Shanghai 201203, China

* Corresponding author:
Jing-Zhe Jiang: jingzhejiang@gmail.com
Ming Duan: duanming@ihb.ac.cn
Dian-Chang Zhang: zhangdch@163.com

[#] These authors contribute equally to this work.

Abstract

Background:

Viruses are the most abundant biological entities, and they play critical roles in entire ecosystems. Nevertheless, current knowledge about them is no more than 1% of the estimated diversity of the Earth's virosphere. Oysters are filter-feeding molluscan bivalves and are ideal sentinels for marine virus exploration and viral ecology studies.

Results:

Here we report a Dataset of Oyster Virome (DOV) that contains 728,784 nonredundant viral operational taxonomic unit (vOTU) contigs and 3,473 high-quality viral genomes, enabling the first comprehensive overview of viral communities in oysters. As in other marine viromes, families Siphoviridae, Podoviridae, and Myoviridae are dominant in the DOV. However, Circoviridae is the most abundant family among the high-quality genomes, indicating that oysters may be their potential hotspots. Despite performing target amplification for RNA genomes, the diversity of RNA viruses was much lower than the diversity of DNA viruses. Notably, most of the vOTUs in the DOV were previously undescribed viruses and could not be clustered with any sequences in three reference datasets. Three approaches (based on references, vOTUs, and auxiliary metabolic genes) consistently showed that host health status, location, and sampling date had potential impacts on virome structures.

Conclusions:

This study highlights the practicality of oysters for marine virus exploration and provides a new direction to understand the relationship between marine bivalves and the environment.

Keywords: *Crassostrea hongkongensis*; Bivalve; Mollusk; Metagenome; Circoviridae; Viral-Like Particle Enrichment; South China; Multiple Displacement Amplification; phi29

1 Background

As the most abundant biological entities on Earth, viruses can hijack organisms from every branch of the tree of life. They play critical roles in host mortality, metabolism, physiology, and evolution, impacting marine biogeochemical cycling and shaping the Earth's microbiomes (Fuhrman, 1999; Wommack and Colwell, 2000; Suttle, 2007). Culture-independent next-generation sequencing technologies have recently been used to explore the tremendous diversity of the virosphere from multiple samples, and as a result, viral genome datasets have expanded exponentially (Paez-Espino et al., 2016; Shi et al., 2016; Shi et al., 2018; Gregory et al., 2019; Wolf et al., 2020; Camarillo-Guerrero et al., 2021; Nayfach et al., 2021). The Integrated Microbial Genome/Virus (IMG/VR) 3.0 system contains more than 18,000 cultivated and 2.31 million uncultivated viral genomes, which were preliminarily clustered into approximately 935,000 viral operational taxonomic units (vOTUs), roughly equivalent to 935,000 viral species (Roux et al., 2021). Although significant progress has been made, these numbers represent only approximately 1% of the conservative estimate of the world's viral diversity (Geoghegan and Holmes, 2017). There is an urgent need to expand the geographical and ecological diversity of the sample collection, and thus decipher the viral "dark matter" to the maximum extent (Krishnamurthy et al., 2017).

Marine animals are teeming with viruses that inhabit the hosts' surfaces, body spaces, and blood (Scanes et al., 2021). These viruses are known as the virome, forming a connection with their host, which is vital to the interaction of the microbe community both in and outside the host's body. Oysters in family Ostreidae are molluscan bivalves and the most highly produced marine shellfishes in the world. China is the largest oyster producer, accounting for 85.3% of the world's total production (FAO, 2019). As keystone species, oysters provide essential benefits to coastal and

estuarine ecosystems, improving water quality and providing a critical habitat for various organisms (Zhang et al., 2012; Powell et al., 2018). Oysters also have the potential to become ideal sentinels for marine virus monitoring and viral ecology studies for the following reasons (Bedford, 1978; Olalemi et al., 2016). First, bivalves, such as oysters and mussels, are sedentary and thus effective in monitoring the environmental conditions for a given area. Second, oysters and mussels are widespread species with large populations, a property that permits frequent sampling, comparisons, and statistical testing of results. Third, and most importantly, bivalves are filter-feeding animals. One oyster can draw up to 5 L of water per hour through their gills and concentrate suspended microbes and particles by factors of a thousand to a hundred thousand times their seawater concentrations. The enrichment of human enteric viruses (Newell et al., 2010) and mimiviruses (Andrade et al., 2015) in oyster gill or gut tissues emphasizes the ability of oysters to accumulate environmental viruses. The United States Mussel Watch Program is the longest-running continuous contaminant monitoring program, which tracks the contamination levels of more than 150 analytes in the environment by monitoring bivalve tissue concentrations (Goldberg, 1986; National Centers for Coastal Ocean Science, 2021); however, this program did not monitor viruses.

The Ostreid Herpesvirus is the first reported and extensively studied pathogen for oysters and many other aquaculture bivalves (De Lorgeril et al., 2018; Gao et al., 2018; Rosani & Venier, 2017; Renault et al., 2000; Farley et al., 1972). Several suspected virus families have been recorded in diseased oysters, including Papovaviridae, Iridoviridae, Togaviridae, Reoviridae, Birnaviridae, and Picornaviridae. Previously, the identification of viral pathogens was based mainly on electron microscopy observations, which were seldom validated by nucleic acid tests or sequencing (Renault and Novoa, 2004). Rosani et al. (2017 and 2019) assembled 26 novel and nearly complete RNA

virus genomes from the public transcriptomic data of *Crassostrea gigas* and *C. corteziensis*, and Zhang et al. (2021) reported four new RNA virus genomes from *C. gigas,* which were recovered from a virome survey of marine invertebrates. Another 33 novel RNA viruses were identified from mixed bivalve samples (including two oyster species *C. hongkongensis* and *C. ariakensis*) (Shi et al., 2016). However, compared with the extensively studied marine virosphere, marine animal-related studies are very few and mainly focused on transcriptomic data (Wolf et al., 2020), thus missing the more important DNA viruses in the marine environment.

Here we report an extensive Dataset of Oyster Virome (DOV) that consists of 54 sequencing libraries from different tissues, sampling sites, and times of *C. hongkongensis,* the largest farming species of oyster along the coast of South China. Using viral-like particle (VLP) enrichment and different genome amplification strategies, we built a knowledge landscape of the oyster's virome community, function, and influencing factors of both RNA and DNA viruses for the first time. This study increased the number of oyster-related virus diversities by tens of thousands of times, identifying that most were previously unrecognized virus categories.

2 Material and methods

2.1 Oyster sampling

The oyster samples collected in this study span five years, from June 2014 to July 2019. We divided the samples into nine **time** batches according to the chronological order. In addition, all the samples were divided into four other groups: four **amplification** groups based on the amplification methods (whole genome amplification (WGA), whole transcriptome amplification (WTA), reverse transcription and WGA (RT-WGA), and double-stranded DNA (dsDNA)); two **tissue** groups based

on tissue origin (mixed tissues and hemolymph of adults); two **status** groups based on health status (diseased and moribund), and seven **Site** groups based on sampling sites (BH, HD, LJ, SZ, TS, YJ, and ZH) (Fig. 1E). In total, we constructed 54 sequencing libraries with 35 samples. For more information, see Table S1.

For eight of the nine time batches, the tissues without the gonad from three oysters were mixed into one sample; the seventh batch was the exception. The **first** batch, dCh, contained samples of dying adult *C. hongkongensis* collected from an oyster farming area in Beihai (BH) of Guangxi Province in June 2014. The **second** batch had two groups, YJd and YJr, and contained samples of healthy adult *C. hongkongensis* collected in September 2015 from an oyster farming area in Yangjiang (YJ), Guangdong. The downstream amplification method for YJd was WGA (to detect mainly DNA virus genomes) and for YJr it was WTA (to detect mainly RNA virus genomes). The **third** batch had eight groups, LJd, LJr, QZd, QZr, TWd, TWr, ZHd, and ZHr, and contained healthy adult *C. hongkongensis* collected from oyster farming areas in the Qinzhou area (QZ) of Beihai (BH), Tanwei area (TW) of Huidong (HD), Zhuhai (ZH), and Lianjiang (LJ) of Guangdong Province in November 2015. The **fourth** batch had two groups, SZd and SZr, and contained healthy adult *C. hongkongensis* collected from the Shenzhen (SZ) oyster farming area in Guangdong in April 2016. (The letters “d” and “r” indicate WGA and WTA, respectively, in the third and fourth batches.) The **fifth** batch, ML, contained healthy adult *C. hongkongensis* collected from SZ in May 2016. The downstream amplification method for ML was RT-WGA (to detect both DNA and RNA virus genomes). The **sixth** batch, BH, contained moribund adult *C. hongkongensis* collected from BH in July 2016. The **seventh** batch had nine groups. GX, K1ZY, K2ZY, T2S, T4S, T5S, T6S, T8S, and ZH, and contained adult *C. hongkongensis* that were separately collected from BH in Guangxi

Province, Kaozhouyang (K#ZY) of Huidong (HD), Taishan (T#S), and Zhuhai (ZH) in Guangdong Province in May 2017. K1ZY, K2ZY, and T8S contained healthy adult *C. hongkongensis*, and the other groups contained moribund adult *C. hongkongensis* . The method of sampling in this batch was different from the method used for all the other batches. A 1-mL syringe was used to draw hemolymph from the pericardial cavity of oysters, and samples from 5–8 oysters were mixed into one sample. The **eighth** batch, os, contained adult *C. gigas* collected in July 2018. The samples in these eight batches were collected and preserved by the South China Sea Fisheries Research Institute (Guangdong, China). The **ninth** batch had two groups, HSd and HSr, and contained healthy adult *C. hongkongensis* purchased from the Huangsha Aquatic Product Market in Guangzhou (GZ), Guangdong, in July 2019; their original farming location was ZH. The samples in this batch were collected and preserved by Guangdong Magigene Technology Co., Ltd (Guangzhou, China). All the samples were quickly frozen in liquid nitrogen, temporarily stored during transportation, and placed in an ultra-low temperature freezer at −80°C for long-term storage.

2.2 VLP enrichment

All 35 samples were processed to enrich for VLPs as described by Wei et al. (2017) and using the online protocols (dx.doi.org/10.17504/protocols.io.m4yc8xw). First, 500 mg of mixed tissue, or 14–34 mg spat mixture, was dissected and ground to powder in liquid nitrogen. The powder was further homogenized in approximately 2–5 volumes of sterile SB buffer (0.2 M NaCl, 50 mM Tris-HCl, 5 mM CaCl2, 5 mM MgCl2, pH 7.5). After three rounds of freezing and thawing, the pellets were resuspended entirely in 10 volumes of pre-cooled SB buffer. For the hemolymph sample, 10 mL hemolymph was mixed with an equal volume of 2×SB buffer, then directly subjected to three rounds

of freezing and thawing. The following steps were the same for the tissue, spat, and hemolymph samples. All the samples were centrifuged at 1,000, 3,000, 5,000, 8,000, 10,000, and 12,000 × g for 5 min each at 4°C using a 3K30 centrifuge (Sigma, Osterode am Harz, Germany), and the supernatants were retained. Cell debris, organelles, and bacterial cells were further removed using a Millex-HV 0.22 μm filter. The filtrates were transferred to ultracentrifuge tubes containing 28% (w/w) sucrose using a syringe. The tubes were transferred to an ice bath for 10 min before centrifugation in a Himac CP 100WX ultracentrifuge (Hitachi, Tokyo, Japan) at 300,000 × g for 2 hr. Supernatants were discarded and the precipitates were fully resuspended in 720 μl of water, 90 μl 10 × DNase I Buffer, 90 μl DNase I (1 U/μl), and incubated at 37°C with shaking for 60 min, followed by storage overnight at 4°C, and transfer to 2-ml centrifuge tubes.

2.3 Viral nucleic acid extraction and amplification

Total nucleic acid was extracted from the VLPs using an HP Viral DNA/RNA Kit (R6873; Omega Bio-Tek, Norcross, USA); carrier RNA was not used, to avoid potential interference with sequencing results. A Qubit™ dsDNA HS Assay Kit (Q32851) and Qubit™ RNA HS Assay Kit (Q32855) (Thermo Fisher Scientific, Waltham, USA) were used to quantify the concentrations of dsDNA and RNA separately.

Virome studies are highly reliant on amplification because the viral biomass in natural samples is very low (Polson et al., 2011; Bar-On et al., 2018). Because most amplification methods introduce bias, it is challenging to study viromic data quantitatively at present (Parras-Moltó et al., 2018; Fan et al., 2021). Here, a REPLI-g Cell WGA and WTA Kit (150052, Qiagen, Hilden, Germany), which is based on the multiple displacement amplification (MDA) method, was used to uniformly amplify

183 the genomes (WGA) and transcriptomes (WTA) (Hosono et al., 2003; Pan et al., 2013; Picher et al.,
184 2016). MDA has many significant advantages over other amplification methods, such as replicating
185 up to 70 kb, more even coverage, and 1000-fold higher fidelity than Taq polymerase amplification
186 (Hosono et al., 2003; Stepanauskas et al., 2017), which make MDA widely used in virome studies.
187 To better compare the RNA and DNA virus communities, we used WGA and WTA methods to
188 construct libraries in four batches of mixed tissues, which accounted for 70% (38/54) of all libraries
189 (Table S1). RT-WGA is a modified protocol that simultaneously amplifies DNA and RNA (Wei et
190 al., 2018b; Li et al., 2019). In this study, 14 libraries were constructed based on RT-WGA, including
191 hemolymph and mixed tissue samples (Table S1). The steps for the WGA, WTA, and RT-WGA were
192 according to the online protocols (dx.doi.org/10.17504/protocols.io.m5vc866). For WTA, there is a
193 "DNA wipeout" step before reverse transcription that aims to remove DNA altogether, but this step
194 is not part of the WGA and RT-WGA protocols. Compared with WTA and RT-WGA, the WGA
195 protocol skips the reverse transcription reaction to avoid amplifying RNA in the downstream
196 reaction. In addition, two other samples were directly subjected to random shotgun library
197 preparation using a Nextera XT DNA Library Preparation Kit (Illumina) following the standard
198 manufacturer's protocol. Because of the limited data quality and sample number, these two libraries
199 were not included in the following diversity analysis.

200 2.4 Library construction and sequencing

201 Amplified DNA was quantified by gel electrophoresis and Nanodrop 2000 spectrophotometer
202 (Thermo Fisher Scientific) and randomly sheared by ultrasound sonication (Covaris M220) to
203 produce fragments ≤800-bp long. The sticky ends were repaired, and adapters were added using T4

DNA polymerase (M4211, Promega, USA), Klenow DNA Polymerase (KP810250, Epicentre), and T4 polynucleotide kinase (EK0031, Thermo Fisher Scientific, USA). Fragments of 300–800 bp were collected after electrophoresis. After amplification, libraries were pooled and subjected to 150 bp, 250 bp, or 300 bp paired-end sequencing on Novaseq 6000, HiSeq X ten, and Miseq platforms (Illumina, USA). Considering the RT-WGA libraries were likely to have higher virus diversity than the WGA and WTA libraries (Wei et al., 2018a), they were sequenced with higher depth and also produced better assembly results (Table S1).

2.5 Virus detection and quantification based on reference viral genomes

Instead of using the traditional read alignment tools such as BLAST, BWA, and Bowtie2, we used FastViromeExplorer (Tithi et al., 2018), which was developed for fast and accurate virus detection and quantification in metagenomics data. FastViromeExplorer filters the alignment results based on minimal coverage criteria and the minimal number of mapped reads and accurately reports virus types and relative abundances. The Kallisto (version 0.43.1) method, integrated with FastViromeExplorer, was used with the default settings to map clean reads against three reference databases: the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database, Global Ocean Virome database (GOV; Roux et al., 2016), and the Integrated Microbial Genome/Virus (IMG/VR) system, separately, to generate a reference abundance table. The RefSeq database contained 14,042 viral genomes or genome segments (update till 30 May 2019), GOV (Roux et al., 2016) includes 298,383 epipelagic and mesopelagic viral contigs, and IMG/VR contains 125,842 metagenomic viral contigs of the set of sequences collected from the Joint Genome

Institute “earth virome” study (Paez-Espino et al., 2016).

2.6 Virus detection and quantification based on *de novo* assembly (vOTU annotation)

High-quality clean reads were generated using Fastp (version 0.20.0) (Chen et al., 2018), (options: --correction, --trim_poly_g, --trim_poly_x, --overrepresentation_analysis, --trim_front1=16, --trim_tail1=2, and --length_required=50) and reads that matched the Illumina sequencing adapters were removed (option: --detect_adapter_for_pe). The clean reads in libraries that were in the same assembly group were pooled and assembled using Megahit (version 1.2.9) (Li et al., 2015) with the default settings. Only contigs longer than 800 bp were kept. To detect low abundant contigs, clean reads that did not map back to the first round of assembled contigs were reassembled for two additional rounds, then all remaining reads were pooled and assembled together. Contigs from all four assembly rounds were pooled, and clustered at 97% global average nucleotide identity with at least 90% overlap of the shorter contig using cd-hit-est (version 4.8.1) (options: -aS 0.9 -c 0.97 -G 1 -M 0 -T 0 -g 1), resulting in 3,347,421 nonredundant contigs (Fig. 1A).

Diamond (Buchfink et al., 2015) is a state-of-the-art method that can annotate sequences with high precision and speed. Compared with BLAST searches against virus only databases, BLAST searches against the NCBI nr database of nonredundant protein sequences can significantly lower the number of false positive results (Nouri et al., 2018; Yao et al., 2020). However, BLAST has relatively low accuracy for short fragments (Jiang et al., 2011) and it cannot be used for sequences that have no similarity. Although, it is possible to dig deeper into the “dark matter” by combining multiple virus mining tools, such as CheckV (Nayfach et al., 2021) and VirSorter2 (Guo et al., 2021),

identifying and classifying suspected viral sequences is challenging because there is a lack of adequate credible annotations (Handley et al., 2019). Given these challenges, the nonredundant contigs were annotated only using Diamond (version 0.9.24.125, options: -e 1e-10, --max-target-seqs 50) against the NCBI nr database (as of 11 July 2019). Among them, 728,784 (21.77%) of the total contigs were annotated as viral origin (i.e., vOTUs). Among them, 7.68% were Eukaryota, 0.34% were Archeae, 21.59% were bacteria, and 0.82% were unclassified cellular organisms, and 47.89% unknown origin (Fig. 1A). FastViromeExplorer was used with default settings to map the clean reads against the vOTU contigs to obtain the vOTUs abundance table.

2.7 Viral genome integrity, taxonomy, and auxiliary metabolic genes analysis

The viral genome completeness of assigned contigs was tested using CheckV (version 0.7.0) and its associated database (Nayfach et al., 2021). After removing false positive contigs that matched more host genes than viral genes, 3,473 nearly complete viral genomes were obtained.

Three methods (Diamond, vContact2, and PhaGCN) were used to determine the taxonomy of the viral contigs at the family level. Diamond annotations were further processed using two scripts (daa2rma and rma2info) in MEGAN6 (Huson et al., 2016) with default parameters, and parsed to taxonomy annotations. The advantage of Diamond is that there is no minimum length requirement for query sequences; however, it has three drawbacks: low accuracy, low annotation rates, and inaccurate taxonomy of NCBI. PhaGCN is a novel semi-supervised learning model that combines the strengths of a BLAST-based model and learning-based model using a knowledge graph (Shang et al., 2017). For comparison purposes, only vOTUs that were longer than 10 kb were analyzed

using PhaGCN and vContact2 with default parameters.

To mine the auxiliary metabolic genes (AMGs) from DOV, Vibrant v1.2.1 (Kieft and Anantharaman, 2020) was used. Salmon v1.5.2 (Patro et al., 2017) was used with default settings to map clean reads against the AMG dataset to obtain the AMGs abundance table.

2.8 Viral contamination assessment

The experimental preparation for viromic sequencing involves the use of various reagents, many of which have been proved to carry contaminated viral sequences of unknown origin (Holmes, 2019). The extent of viral contamination in common laboratory components, especially viruses with small single-stranded DNA (ssDNA) genomes, has been reported previously (Asplund et al., 2019; Ashleigh et al., 2021).

To assess the viral contaminant level in this study, all the 3,347,421 nonredundant contigs (≥800 bp; not only viral contigs) in DOV were used as queries in a BLASTN search (with the parameters set as 95% identity and 95% query coverage) against the approximately 500 contaminant viral sequences reported by Asplund et al. (2019) and Ashleigh et al. (2021). We found very little evidence of viral contamination; no sequences matched with 100% identity, no expected circoviruses or RNA viruses were detected, and most of the alignments were with dsDNA phages (Additional file 1). The 3,473 near-complete viral genomes were used as queries in the same BLASTN search, but no matches were found. We also used Salmon (v1.5.2) to map all the clean reads in the DOV libraries to the contaminant viral sequences. The mapping rates for most of these libraries were <0.01% (Additional file 2), which is consistent with the BLASTN results.

2.9 Viral community and statistical analysis

In this study, the FPKM (Fragments Per Kilobase per Million) value was used to represent the relative abundance of the reference viral genomes, vOTUs, and AMGs. On the basis of the FPKM-transformed abundance table, R and Excel were used to analyze the corresponding viral diversity and community structures. The vegan and ggplots R packages were used to calculate α-diversity indexes and plot the nonmetric multidimensional scaling (NMDS). Analysis of variance (ANOVA) and TukeyHSD were used to test the differences between groups with the significance level set at 0.05. For the procrustes analysis, the characteristic axis coordinates of NMDS were extracted as the input of the procrustes() function, and the protest() function was used to perform the substitution test to evaluate the significance of the results.

For core species analysis, vOTUs with >10% relative abundance and detected in >50% of the libraries were defined as core species, and the top 5000 abundant vOTUs were taken as the nodes for co-correlation network analysis. The correlation between nodes was calculated by Spearman's rank correlation coefficient; the threshold of correlation coefficient (r) was set at 0.7 and the significance parameter (p) was set at 0.01. According to the constructed network graph file, the topological attributes (degree and closeness centrality) of the corresponding nodes are calculated. Based on topological properties, the core vOTU in the viral community were inferred. The Fruchterman Reingold layout in Gephi (version 0.9.2) was used to visualize the network.

3 Result and discussion

3.1 Overview of the Dataset of Oyster Virome (DOV)

For this study, we collected 35 samples of mixed tissue or hemolymph from *C. hongkongensis* at nine time points from seven major oyster farming areas along the coast of South China (Fig. 1, Table S1). Fifty-four oyster virome libraries were constructed using three primary amplification methods (WTA, WGA, and RT-WGA) and sequenced (Table S1). A total of 3,347,421 nonredundant contigs (≥800 bp) were obtained after assembly. Among them, 728,784 (21.77%) were annotated as viral origin by an alignment-based method (Fig. 1A), which we called the DOV. The viral contigs were assembled mainly from the RT-WGA libraries of hemolymph samples with higher sequencing coverages (Fig. 1B). Rarefaction curves (Fig. 1C) show that the sequencing depths were sufficient, and the vOTU numbers in the WTA libraries were the lowest among all the libraries.

Notably, the ratio of viral reads (mapping rate) varied a lot depending on the reference databases that were searched (Fig. 1E). The mapping rate of *de novo* assembled vOTUs (29.81%) was much higher than the mapping rates of the RefSeq (NCBI viral reference genomes) (3.50%) and the RefSeq plus two other public virus datasets (GOV and IMG/VR) (12.06%) (Fig. 1E, Table S1). The high mapping rates of vOTUs confirmed that the VLPs enrichment protocol was effective (Wei et al., 2018b; Liu et al., 2019), and filter-feeding oysters can efficiently accumulate environmental particles, including largely unknown viruses (Bedford et al., 1978; Olalemi et al., 2016). To our knowledge, this is the biggest viral metagenomic dataset for any marine animal currently available.

3.2 Taxonomy of DOV

Viruses in order Caudovirales dominated the oyster virome (Figs. 2, S1), just as they dominate the public dataset and culture collections (Kauffman et al., 2018). Siphoviridae (28.5%–30.61%), Podoviridae (13.46%–42.52%), and Myoviridae (18.36%–29.61%) were the top three Caudovirales families in the DOV (Fig. 2A–C). Because of a primary bias of MDA, circular ssDNA viruses (including Microviridea and Circoviridea) accounted for only 2.23% of all the viruses (Fig. S1), which means their diversity is much lower than the diversity of the dsDNA viruses in DOV. Consistent with the rarefaction curves (Fig. 1C), RNA viruses accounted for only 0.68% (4,958 vOTUs) of all the viruses in DOV (Figs. 1D, S1), and DNA viruses dominated at all the sampling sites (Fig. 1D).

Diamond annotation of short contigs has limited accuracy (Jiang et al., 2011), and a large proportion of them (34.88%) could not be assigned at the family level (Fig. S1). For the long contigs (≥10 kb), PhaGCN successfully classified 6,362 out of 8,760 vOTUs (Fig. 2B), which far exceeded the number classified by vContact2 (214/8,760) (Fig. 2C). Among them, Siphoviridae, Podoviridae, and Myoviridae accounted for 31,12%, 23.37%, and 24.81% of the vOTUs, respectively, and the unassigned vOTUs decreased to 11.46% (Fig. 2D). vContact2 clusters were consistent with the color scheme of the PhaGCN families and their sources, which further proved the consistency of the two results (Fig. 2D). Impressively, the DOV nodes accounted for 74.58% of the with vOTUs, whereas the RefSeq nodes account for only 25.42% of the RefSeq viral genomes in the vContact2 network (Fig. 2E), indicating that current knowledge about the ocean virosphere is insufficient. Oysters and other filter-feeding bivalves may act as viral hotspots and play invisible but essential roles in regulating the marine microbiome.

3.3 Near-complete viral genomes

The integrity of the virus genomes was evaluated using CheckV (Nayfach et al., 2021). A total of 3,473 viral contigs with >90% genomic completeness (including full-length genomes) were identified (Figs. 3, S2, Table S2). Their genome lengths ranged from 1,206–60,277 bp, and the GC content ranged from 24.74%–65.70% (Fig. 3). Their identity with known viral proteins was generally low (0%–40%) (Fig. S2), indicating that most of the contigs belonged to new viral categories. Only 16 of the genomes clustered with nonredundant reference genomes using CheckV at 95% average nucleotide identity and 70% alignment fraction of contig. Therefore, we considered that unclassified and unknown viruses accounting for 66.9% (2,369) of the genomes at the family classification level (Table S2).

The classified genomes belonged to at least 11 DNA virus families. Those in order Caudovirales included Podoviridae (45), Sipoviridae (43), Myoviridae (14), and Autograhpiviridae (8) (Fig. S2). Circoviridae (order Cirlivirales) and Microviridae (order Petitvirales) were the most abundant families, accounting for 11.27% (399) and 6.98% (247) of the classified genomes, respectively (Fig. S2, Table S2). Moreover, in the viral proteomic phylogenetic tree, Circoviridae branches were widely dispersed and mixed with unannotated branches (Fig. 3), implying that many putative circovirus clades are yet to be identified. All known hosts of circoviruses are in clade Bilateria in kingdom Animalia (Virus-Host Database, May 2021). Whether these circoviruses are pathogens or live as symbionts in oyster hosts, and their diversity and evolutionary position in primitive bivalves need further study.

3.4 Comparisons among amplification strategies

MDA introduces bias by prioritizing circular ssDNA genome (Binga et al., 2008), and this may have led to the >80% abundance of circular ssDNA virus in several libraries in this study (Fig. S3). Parras-Moltó et al. (2018) found that ordination plots based on dissimilarities among vOTU profiles showed perfect overlapping of related amplified and unamplified viromes and strong separation from unrelated viromes, which showed that MDA can be used for community studies. In this study, WGA and WTA amplified libraries were scattered into different clusters (Fig. 4A–C). Because RT-WGA simultaneously amplifies both DNA and RNA genomes (Fig. S3), the RT-WGA libraries overlapped with the WGA and WTA libraries in the NMDS plot (Fig. 4A). However, the higher variation of richness and Shannon indexes for the RT-WGA libraries (Fig. S4A-C) implies a corresponding increase in the community's complexity.

Different MDA strategies can efficiently target different genomes, because the vOTUs of RNA viruses in the WTA libraries significantly outnumber those in the WGA libraries, and vice versa for the DNA viruses (Fig. S4D, E). Although the differences in α-diversity indexes were not very significant, similar rules still exist (Fig. S4A–C), which is consistent with previous obversions (Figs. 1C, S1). It seems to be common that the diversity of DNA viruses in nature and public databases is higher than the diversity of RNA viruses (Roux et al., 2021; Rosario et al., 2018; Levin et al., 2017). The extremely high mutation rates of RNA genomes challenged the accuracy of alignment-based detections (Holmes, E. C., 2009; Shi et al., 2016). The instability of RNA genomes and potential amplification bias further complicated the comparisons among amplification strategies. However, quantitively comparing the diversity and abundance among RNA and DNA viruses in the real world will be very interesting (Holmes, E. C., 2011; Zhu et al., 2021; Steward et al., 2013).

3.5 Influences on the viral community

We further evaluated the correlation among various community parameters, including the quantity and quality of sequencing reads, the vOTU counts, the ratio of viral reads or host reads, and the diversity indexes (Fig. 4D). The α- or β-diversities correlated well between the reference-based and the vOTU-based virus detection methods (Fig. 4B, D), which indicates that both community-deciphering approaches apply to this study. Although the reference-based method was not sensitive enough, it can still provide basic information about communities as long as the reference dataset is large and relevant. Besides amplification, differences between tissue groups were also prominent (Fig. 4A; higher F-value), but they are not discussed here because the tissue was not the only variable in the designed cohort batches and groups (Table S1). Rather, this study focused on the influences of three factors, health status, sampling site, and sampling date, on communities in parallel cohorts (Fig. 4).

Diverse communities have a higher ability to resist invasion by exotic species than simple communities (Levine & D'Antonio 1999). A decrease in the α-diversity of a microbial community is usually associated with disease in the host, including human, mouse, and some marine invertebrates (shrimp, oyster, and sea cucumber) (Petersen and Round, 2014; Sandra et al., 2020). However, we did not detect the expected differences in α- or β-diversity between moribund and healthy groups in any of the libraries (Fig. S5A) or in the parallel cohorts (Fig. S5B). Nevertheless, we did observe consistent correlations between α-diversity and virus abundance (mapping ratio of viral reads) in many groups, including the WGA, WTA, tissue, and all libraries (Fig. S6). The high abundance and low diversity imply that one or a few viruses dominate the host, which is associated with infection (Shi et al., 2016; Zinter et al., 2019; Chiu et al., 2019). Given this, the α-diversity of

410 the virome could be a feasible criterion for demonstrating a causal link between the presence of a
411 virus and the onset of disease in bivalve mollusks.

412 Geographical origin (site) also substantially influenced the community. Samples from the same
413 location tended to aggregate, and significant differences in α-diversity were observed in the WGA
414 and WTA groups (Figure S7). The influence of the habitat on the microbiome of the host has been
415 reported in many animals (Ge et al., 2021; Krotman et al., 2020; Sandri et al., 2020; Su et al., 2020),
416 and environmental variations may be one of the reasons for differences in α-diversity (Oetama et
417 al., 2016). However, unlike freely swimming fish, oysters are sedentary and filter large volumes of
418 the surrounding water daily (Bedford, 1978; Olalemi et al., 2016). The influence of site on the
419 viromic community was weaker than that of the time point (Fig. 4A), and this was reflected in the
420 proportion of unique vOTUs (i.e., those that were detected only in one group) (Fig. S8). The
421 relatively high proportion of unique vOTUs in the time period groups implies that viral communities
422 are dynamic with time, and the low proportion of unique vOTUs indicates that viruses actively
423 communicate among locations. However, because of the limited sample number and the diversity
424 of host species, these results need further verification.

425 3.6 Core species and AMG diversity

426 Core species are an essential concept in ecology and play critical roles in maintaining community
427 structure (Berg et al., 2020). However, it is not clear whether core species exist in viral communities.
428 Here we defined core species according to conventional standards and found two different patterns
429 in oyster viromes. For the DNA virus community, the pattern is quite similar to that of the
430 microbiome. The abundance of core species was close to 80% in the WGA libraries, but they

accounted for 20% or less of the total vOTU counts (Fig. 5A, B). For the RNA virus community, the richness was low, and therefore the abundance of core species was close to 100% in some libraries (Fig. 5A, B). The co-occurrence network further proves that the core species also occupied core positions in the network (Fig. 5C). Furthermore, their degree (Fig. 5E) and closeness centrality (Fig. 5F) were significantly higher than those of other viral species. These results suggest that the viral community has a core differentiation mechanism similar to that of cellular microorganisms. Whether the core species in the oysters are independent or a reflection of the core species in the water environment is an interesting question.

Viruses not only directly regulate the microbial community by killing their hosts but also play essential roles in metabolic regulation in the marine ecosystem (Suttle, 2005; Breitbart et al., 2012; Breitbart et al., 2018). Of the AMGs, 9,091 were assigned to 12 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, and 98 pathways were identified among the viruses in the DOV (Table S3). Among them, pathways associated with the metabolism of cofactors and vitamins, amino acids, energy, and carbohydrates were significantly enriched (Fig. S9A), which is similar to the results obtained for other marine viromes (Castelán-Sánchez et al., 2020; Hurwitz et al., 2016; Hurwitz et al., 2013). The AMG community (Fig. S9B) showed consistency with the vOTU community (Fig. S9C), and the richness and Shannon index showed positive correlations between the two communities (Figs. 4D, S9D, S9E). These findings indicate that the oyster viromic function was closely related to that of the species community. However, whether the function determines the community or the community determines the function cannot be answered in this study. Besides, the previous finding that viruses with large genomes tend to encode more AMGs than viruses with small genomes and provide ecological functions beyond sustaining basic infection and proliferation

(Jiang et al., 2020) is supported by the results shown in Fig. S9F.

4 Conclusions

The DOV is a comprehensive viromic dataset with high resolution that records the vast diversity and uniqueness of viruses associated with filter-feeding bivalves. This viromic dataset highlights the potential of studying marine animals to discover the hidden marine virosphere of either viral pathogens of the host or environmental viruses ingested and stored by the host. Although further work is need to identify the novel viruses, including circoviruses, RNA viruses, and potential oyster pathogens, this study provides a new perspective for studying and understanding the virome community and its function in marine animals.

5 Funding

This project was supported by the Natural Science Foundation of China (nos. 31972847 and 32172955) to Jiang JZ and Duan M; Financial Fund of the Ministry of Agriculture and Rural Affairs, P. R. of China (NHYYSWZZZYKZX2020) to Zhang DC; the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (nos. 2020TD42 and 2021SD05) to Jiang JZ; the Guangdong Provincial Special Fund for Modern Agriculture Industry Technology Innovation Teams (no. 2019KJ141) to Jiang JZ. The funders had no role in the study design, data collection and analysis, decision to publish, or manuscript preparation.

6 Acknowledgment

We thank Margaret Biswas, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn/) for editing the English text of a draft of this manuscript. We are also very grateful to Dr. Edward C. Holmes, Dr. Curtis A. Suttle, and Dr. Xu Kevin Zhong for their insightful comments and feedback.

7 Ethics approval and consent to participate

Not applicable.

8 Consent for publication

Not applicable.

9 Competing interests

The authors declare that they have no competing interests.

10 Author details

[a] Key Laboratory of South China Sea Fishery Resources Exploitation and Utilization, Ministry of Agriculture, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510300, Guangdong, China. [b] State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, Hubei, China. [c] Guangdong Magigene Biotechnology Co., Ltd, Guangzhou 510000, Guangdong, China [d] School of Medicine, Sun Yat-sen University, Shenzhen 518107, Guangdong, China. [e] Tianjin Agricultural University, Tianjin 300384, China. [f] Shenzhen Fisheries Development Research Center, Shenzhen 518067, Guangdong, China. [g] Bureau of Agriculture and Rural Affairs of Conghua District, Guangzhou 510925, Guangdong, China. [h] Current address: Shanghai Majorbio Bio-Pharm Technology Co., Ltd, Shanghai 201203, China.

11 Authors' contributions

JJZ, DM, and ZDC conceived the study and directed the project; JJZ designed the experiments; JJZ, WHY, SSK, WM, YLT, YT, and LJ obtained the samples; WHY and JJZ conducted the wet lab experiments; JJZ, FYF, and WHY performed data analyses with supports from SM, YLL, GYX, and JT; JJZ and FYF interpreted and visualized the results; JJZ wrote the manuscript with supports from FYF and JT. All authors read and approved the final manuscript.

12 Availability of data and materials

The data set supporting the results of this article has been deposited in the Genome Sequence Archive (GSA) under BioProject accession code PRJCA007058[https://ngdc.cncb.ac.cn/gsub/submit/bioproject/subPRO010366/overview].

13 References

1. Andrade KR, Boratto PP, Rodrigues FP, Silva LC, Dornas FP, Pilotto MR et al. Oysters as hot spots for mimivirus isolation. Arch Virol. 2015;160(2):477-82.
2. Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, Herrera JA, et al.

Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin Microbiol Infect. 2019;25(10):1277-85.
3. Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth. Proc Natl Acad Sci U S A. 2018;115(25):6506-6511.
4. Bedford AJ, Williams G, Bellamy AR. Virus accumulation by the rock oyster Crassostrea glomerata. Appl Environ Microbiol. 1978;35(6):1012-8.
5. Berg G, Rybakova D, Fischer D, Cernava T, Vergès MC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome. 2020;8:103. https://doi.org/10.1186/s40168-020-00875-0
6. Binga EK, Lasken RS, Neufeld JD. Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J. 2008;2(3):233–241.
7. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018;3(7):754-766. https://doi.org/10.1038/s41564-018-0166-y
8. Breitbart M. Marine viruses: truth or dare. Ann Rev Mar Sci. 2012;4:425-48.
9. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184(4):1098-1109.e9.
10. Castelán-Sánchez HG, Meza-Rodríguez PM, Carrillo E, Ríos-Vázquez DI, Liñan-Torres A, Batista-García RA, et al. The microbial composition in circumneutral thermal springs from Chignahuapan, Puebla, Mexico reveals the presence of particular sulfur-oxidizing bacterial and viral communities. Microorganisms. 2020;8(11):1677. https://doi.org/10.3390/microorganisms8111677
11. Chiu CY, Miller SA. Clinical metagenomics. Nat Rev Genet. 2019;20(6):341-355. https://doi.org/10.1038/s41576-019-0113-7
12. de Lorgeril J, Lucasson A, Petton B, Toulza E, Montagnani C, Clerissi C, et al. Immune-suppression by OsHV-1 viral infection causes fatal bacteraemia in Pacific oysters. Nat Commun. 2018;9(1):4215.
13. Fan X, Yang C, Li W, Bai X, Zhou X, Xie H, et al. SMOOTH-seq: single-cell genome sequencing of human cells on a third-generation sequencing platform. Genome Biol. 2021;22(1):195.
14. Farley CA, Banfield WG, Kasnic G Jr, Foster WS. Oyster herpes-type virus. Science. 1972;178(4062):759-60.
15. Gao F, Jiang JZ, Wang JY, Wei HY. Real-time quantitative isothermal detection of Ostreid herpesvirus-1 DNA in Scapharca subcrenata using recombinase polymerase amplification. J Virol Methods. 2018;255:71-75.
16. Ge Y, Jing Z, Diao Q, He JZ, Liu YJ. Host species and geography differentiate honeybee gut bacterial communities by changing the relative contribution of community assembly processes. mBio. 2021;12(3):e0075121.
17. Geoghegan JL, Holmes EC. Predicting virus emergence amid evolutionary noise. Open Biol. 2017;7(10):170189. http://doi.org/10.1098/rsob.170189
18. Goldberg E D. The mussel watch concept. Environ Monit Assess. 1986; 7(1):91-103.
19. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA viral macro- and microdiversity from Pole to Pole. Cell. 2019;177(5):1109-1123. https://doi.org/10.1016/j.cell.2019.03.040
20. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2:

549 a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses.
550 Microbiome. 2021;9(1):37. https://doi.org/10.1186/s40168-020-00990-y
551 21. Handley SA, Virgin HW. Drowning in Viruses. Cell. 2019;177(5):1084-1085.
552 22. Holmes EC. Reagent contamination in viromics: all that glitters is not gold. Clin Microbiol
553 Infect. 2019;25(10):1167-1168.
554 23. Holmes EC. The Evolution and Emergence of RNA Viruses. Oxford Univ Press. 2009.
555 24. Holmes EC. What does virus evolution tell us about virus origins? J Virol. 2011;85(11):5247-
556 51.
557 25. Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, et al. Unbiased whole-genome
558 amplification directly from clinical samples. Genome Res. 2013;13(5):954-64.
559 26. Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and
560 dark ocean. Genome Biol. 2013;14(11):R123. https://doi.org/10.1186/gb-2013-14-11-r123
561 27. Hurwitz BL, U'Ren JM. Viral metabolic reprogramming in marine ecosystems. Curr Opin
562 Microbiol. 2016;31:161-168.
563 28. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community
564 Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. PLoS
565 Comput Biol. 2016;12(6): e1004957.
566 29. Jiang JZ, Zhang W, Guo ZX, Cai CC, Su YL, Wang RX, et al. Functional annotation of an
567 expressed sequence tag library from Haliotis diversicolor and analysis of its plant-like
568 sequences. Mar Genomics. 2011;4(3):189-96.
569 30. Jiang T, Guo C, Wang M, Wang M, Zhang X, Liu Y, et al. Genome Analysis of Two Novel
570 Synechococcus Phages That Lack Common Auxiliary Metabolic Genes: Possible Reasons and
571 Ecological Insights by Comparative Analysis of Cyanomyoviruses. Viruses. 2020;12(8):800.
572 https://doi.org/10.3390/v12080800
573 31. Kauffman KM, Hussain FA, Yang J, Arevalo P, Brown JM, Chang WK, et al. A major lineage
574 of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature.
575 2018;554(7690):118-122. https://doi.org/10.1038/nature25474
576 32. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation
577 of microbial viruses, and evaluation of viral community function from genomic sequences.
578 Microbiome. 2020;8(1):90. https://doi.org/10.1186/s40168-020-00867-0
579 33. Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res.
580 2017;239:136-142.
581 34. Krotman Y, Yergaliyev TM, Alexander Shani R, Avrahami Y, Szitenberg A. Dissecting the
582 factors shaping fish skin microbiomes in a heterogeneous inland water system. Microbiome.
583 2020;8(1):9.
584 35. Levin RA, Voolstra CR, Weynberg KD, van Oppen MJ. Evidence for a role of viruses in the
585 thermal sensitivity of coral photosymbionts. ISME J. 2017;11(3):808-812.
586 https://doi.org/10.1038/ismej.2016.154
587 36. Levine JM, D'Antonio CM. Elton revisited: a review of evidence linking diversity and
588 invasibility. Oikos. 1999;87(1):15-26.
589 37. Li Y, Fu X, Ma J, Zhang J, Hu Y, Dong W, et al. Altered respiratory virome and serum cytokine
590 profile associated with recurrent respiratory tract infections in children. Nat Commun.
591 2019;10(1):2288.
592 38. Liu P, Chen W, Chen JP. Viral metagenomics revealed Sendai virus and Coronavirus infection

of Malayan pangolins (Manis javanica). Viruses. 2019;11(11):979.
39. National Centers for Coastal Ocean Science, 2021: National Status and Trends: Mussel Watch Program, https://www.fisheries.noaa.gov/inport/item/39400.
40. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5):578-585.
41. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6(7):960-970. https://doi.org/10.1038/s41564-021-00928-6
42. Newell DG, Koopmans M, Verhoef L, Duizer E, Aidara-Kane A, Sprong H, et al. Food-borne diseases - the challenges of 20 years ago still persist while new ones continue to emerge. Int J Food Microbiol. 2010;139(Suppl 1):S3-15.
43. Nouri S, Matsumura EE, Kuo YW, Falk BW. Insect-specific viruses: from discovery to potential translational applications. Curr Opin Virol. 2018;33:33-41.
44. Oetama VSP, Hennersdorf P, Abdul-Aziz MA, Mrotzek G, Haryanti H, Saluz HP. Microbiome analysis and detection of pathogenic bacteria of Penaeus monodon from Jakarta Bay and Bali. Mar Pollut Bull. 2016;110(2):718-725.
45. Olalemi A, Baker-Austin C, Ebdon J, Taylor H. Bioaccumulation and persistence of faecal bacterial and viral indicators in Mytilus edulis and Crassostrea gigas. Int J Hyg Environ Health. 2016;219(7 Pt A):592-598.
46. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536(7617):425-30. https://doi.org/10.1038/nature19094
47. Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, et al. Two methods for full-length RNA sequencing for low quantities of cells and single cells. Proc Natl Acad Sci U S A. 2013;110(2):594-9.
48. Parras-Moltó M, Rodríguez-Galet A, Suárez-Rodríguez P, López-Bueno A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. Microbiome. 2018;6(1):119.
49. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-419.
50. Petersen C, Round JL. Defining dysbiosis and its influence on host immunity and disease. Cellular Microbiology. 2014;16(7):1024-1033.
51. Picher ÁJ, Budeus B, Wafzig O, Krüger C, García-Gómez S, Martínez-Jiménez MI, et al. TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. Nat Commun. 2016;7:13296.
52. Polson SW, Wilhelm SW, Wommack KE. Unraveling the viral tapestry (from inside the capsid out). ISME J. 2011;5(2):165-8.
53. Porter AF, Cobbin J, Li C, Eden JS, Holmes EC. Metagenomic identification of viral sequences in laboratory reagents. bioRxiv. 2021. https://doi.org/10.1101/2021.09.10.459871
54. Powell D, Subramanian S, Suwansa-Ard S, Zhao M, O'Connor W, Raftos D, et al. The genome of the oyster Saccostrea offers insight into the environmental resilience of bivalves. DNA Res. 2018;25(6):655-665.
55. Renault T, Le Deuff RM, Chollet B, Cochennec N, Gérard A. Concomitant herpes-like virus

infections in hatchery-reared larvae and nursery-cultured spat Crassostrea gigas and Ostrea edulis. Dis Aquat Organ. 2000;42(3):173-183.
56. Renault T, Novoa B. Viruses infecting bivalve molluscs. Aquat Living Resour. 2004;17(4):397-409.
57. Rosani U, Gerdol M. A bioinformatics approach reveals seven nearly-complete RNA-virus genomes in bivalve RNA-seq data. Virus Res. 2017;239:33-42.
58. Rosani U, Shapiro M, Venier P, Allam B. A needle in a haystack: tracing bivalve-associated viruses in high-throughput transcriptomic data. Viruses. 2019;11(3):205.
59. Rosani U, Venier P. Oyster RNA-seq Data Support the Development of Malacoherpesviridae Genomics. Front Microbiol. 2017;8:1515.
60. Rosario R, Fierer N, Miller S, Luongo J, Breitbart M. Diversity of DNA and RNA viruses in indoor air as assessed via metagenomic sequencing. Environ Sci Technol. 2018;52(3):1014-1027.
61. Roux S, Páez-Espino D, Chen IA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. 2021;49(D1):D764–D775. https://doi.org/10.1093/nar/gkaa946
62. Sandra IV, Roger H, Dean R. J. Microbiome diversity and dysbiosis in aquaculture. Reviews in Aquaculture. 2020;12(2):1077-1096.
63. Sandri C, Correa F, Spiezio C, Trevisi P, Luise D, Modesto M, et al. Fecal microbiota characterization of seychelles giant tortoises (Aldabrachelys gigantea) living in both wild and controlled environments. Front Microbiol. 2020;11:569249.
64. Scanes E, Parker LM, Seymour JR, Siboni N, King WL, Danckert NP, et al. Climate change alters the haemolymph microbiome of oysters. Mar Pollut Bull. 2021;164:111991.
65. Shang J, Jiang J, Sun Y. Bacteriophage classification for assembled contigs using graph convolutional network. Bioinformatics. 2021;37(Suppl_1):i25-i33.
66. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, et al. The evolutionary history of vertebrate RNA viruses. Nature. 2018;556(7700):197-202. https://doi.org/10.1038/s41586-018-0012-7
67. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540(7634):539-543. https://doi.org/10.1038/nature20167
68. Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. Nat Commun. 2017;8(1):84.
69. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we missing half of the viruses in the ocean? ISME J. 2013;7(3):672-9. https://doi.org/10.1038/ismej.2012.121
70. Su S, Munganga BP, Du F, Yu J, Li J, Yu F, et al. Relationship between the fatty acid profiles and gut bacterial communities of the chinese mitten crab (Eriocheir sinensis) from ecologically different habitats. Front Microbiol. 2020;11:565267.
71. Suttle CA. Viruses in the sea. Nature. 2005;437(7057):356-61.
72. Wei HY, Huang S, Wang JY, Gao F, Jiang JZ. Comparison of methods for library construction and short read annotation of shellfish viral metagenomes. Genes Genom. 2018a;40(3):281–288. https://doi.org/10.1007/s13258-017-0629-1
73. Wei HY, Huang S, Yao T, Gao F, Jiang JZ, Wang JY. Detection of viruses in abalone tissue using metagenomics technology. Aquaculture Research. 2018b;49(8):2704-2713.

74. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol. 2020;5(10):1262-1270.
75. Yao Z, Zou C, Peng N, Zhu Y, Bao Y, Zhou Q, et al. Virome identification and characterization of Fusarium sacchari and F. andiyazi: causative agents of Pokkah Boeng disease in sugarcane. Front Microbiol. 2020;11:240.
76. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 2012;490(7418):49-54. https://doi.org/10.1038/nature11413
77. Zhang YY, Chen Y, Wei X, Cui J. Viromes in marine ecosystems reveal remarkable invertebrate RNA virus diversity. Sci China Life Sci. 2021. https://doi.org/10.1007/s11427-020-1936-2
78. Zinter MS, Dvorak CC, Mayday MY, Iwanaga K, Ly NP, McGarry ME, et al. Pulmonary metagenomic sequencing suggests missed infections in immunocompromised children. Clin Infect Dis. 2019;68(11):1847-1855.
79. Zuo T, Liu Q, Zhang F, Yeoh YK, Wan Y, Zhan H, et al. Temporal landscape of human gut RNA and DNA virome in SARS-CoV-2 infection and severity. Microbiome. 2021;9(1):91.

Figures

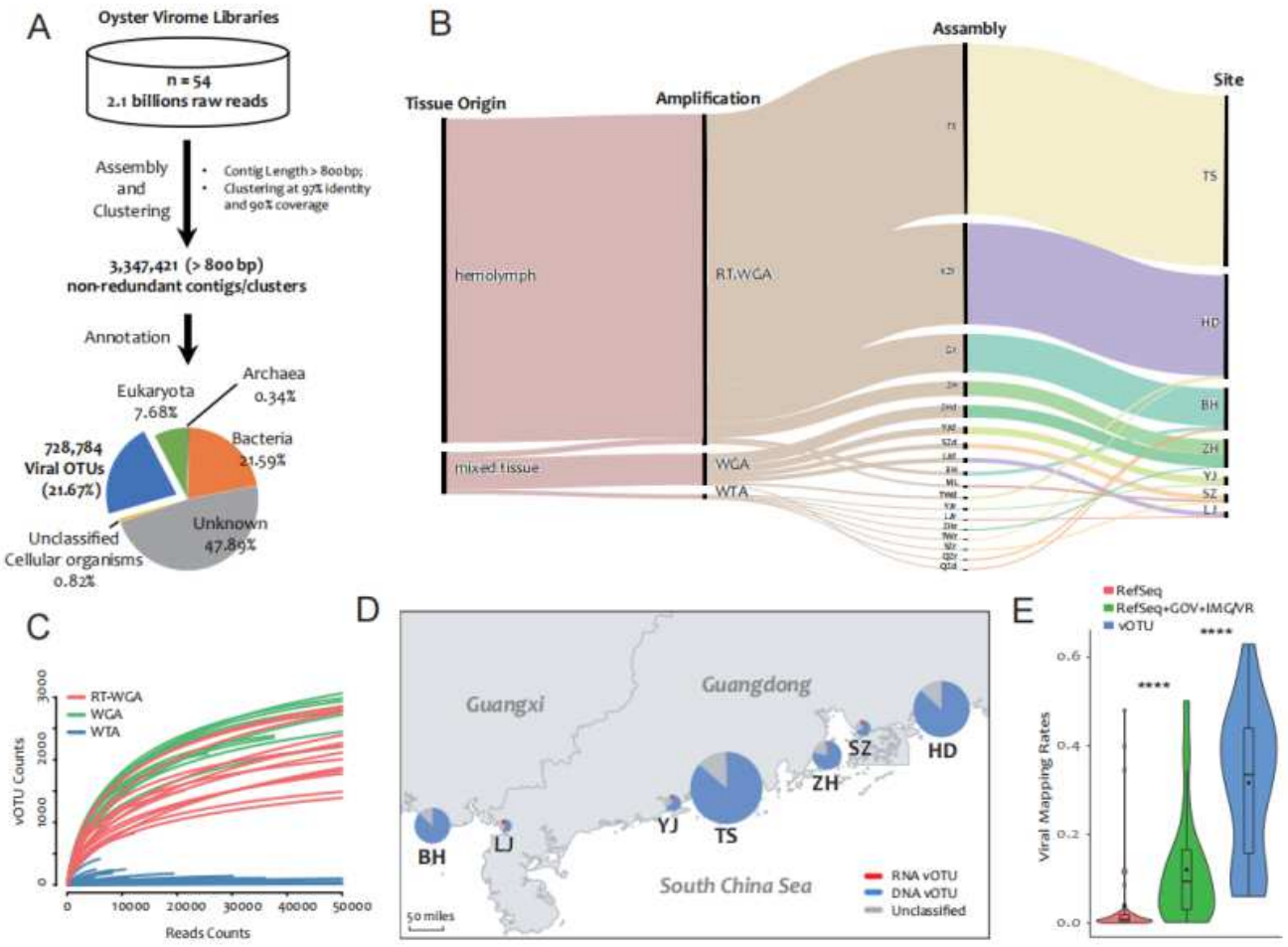

Figure 1

Overview of the Dataset of Oyster Virome (DOV). (A) De novo assembly and annotation pipeline. (B) Sankey diagram of the relationship among different batches and groups. The height of the black vertical bars proportionally represent the number of viral contigs (vOTUs) assembled under teach group. (C) Rarefaction curves of the oyster viromic libraries. RT-WGA, reverse transcription and whole genome amplification; WGA, whole genome amplification; WTA whole transcriptome amplification. (D) Sampling site distribution map and the number of detected vOTUs from each site. The radius of the pie chart indicates the number of DNA, RNA, and unclassified vOTUs. (E) Mapping rates of viral reads in total clean reads. RefSeq, NCBI viral RefSeq genomes (release June 2019); GOV (release July 2020), Global Ocean Virome dataset; IMG/VR (release January 2018), a database of cultured and uncultured DNA viruses and retroviruses maintained by the Joint Genome Institute; vOTU, de novo assembled vOTUs in the DOV. **** indicates p <0.0001 (Student's t-test between the three mapping rates).

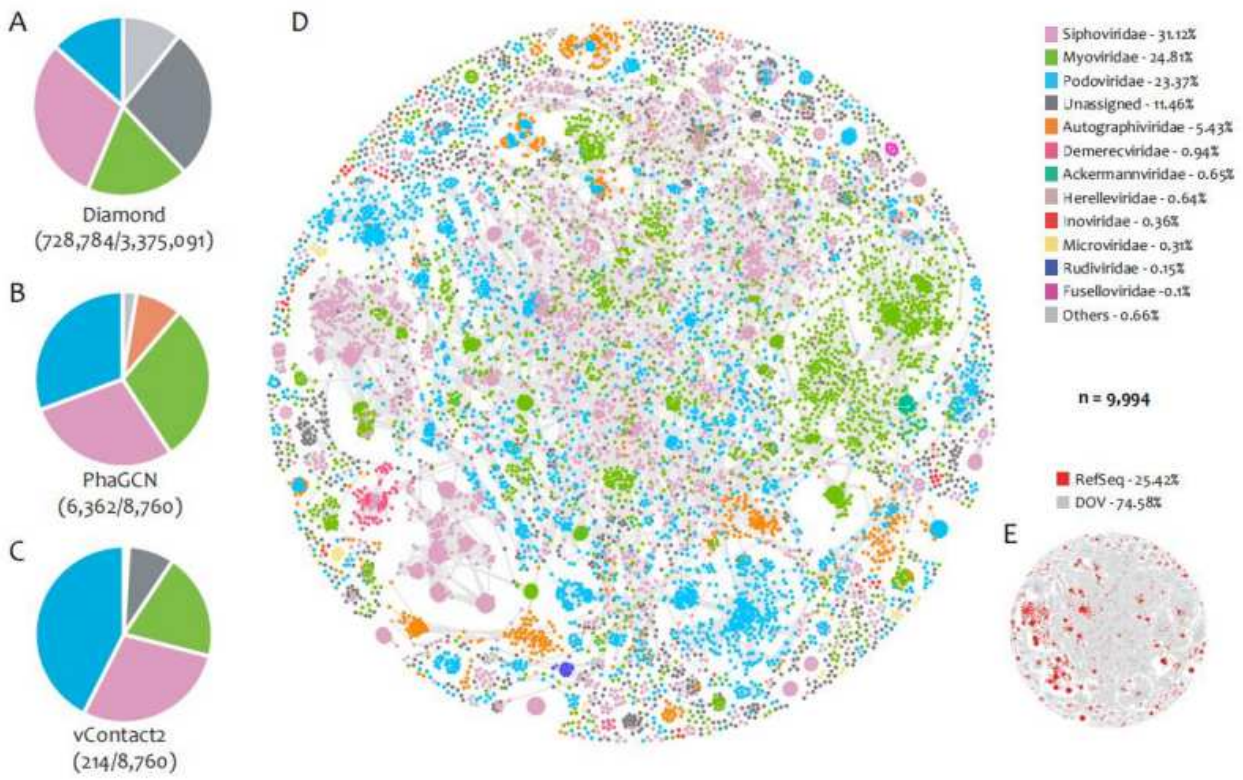

Figure 2

Taxonomy classification of the Dataset of Oyster Virome (DOV) at the family level. (A–C) Pie charts showing the proportion of different viral families in the total viral contigs (vOTUs) longer than 800 bp. The vOTUs were classified using Diamond (v0.9.14) (A), and vOTUs longer than 10 kb were classified using PhaGCN (B) and vContact2 (C). The numbers in parentheses indicate the number of vOTU successfully classified/total number of vOTU. (D, E) vContact2 networks constructed with vOTUs and NCBI RefSeq viral genomes (release June 2019) longer than 10 kb showing they have the same topology. The colors of the nodes indicate different PhaGCN families (D), and their sources (E). n, total number of nodes in vContact2 networks. The percentage of each family or source in (D) is listed after corresponding legends.

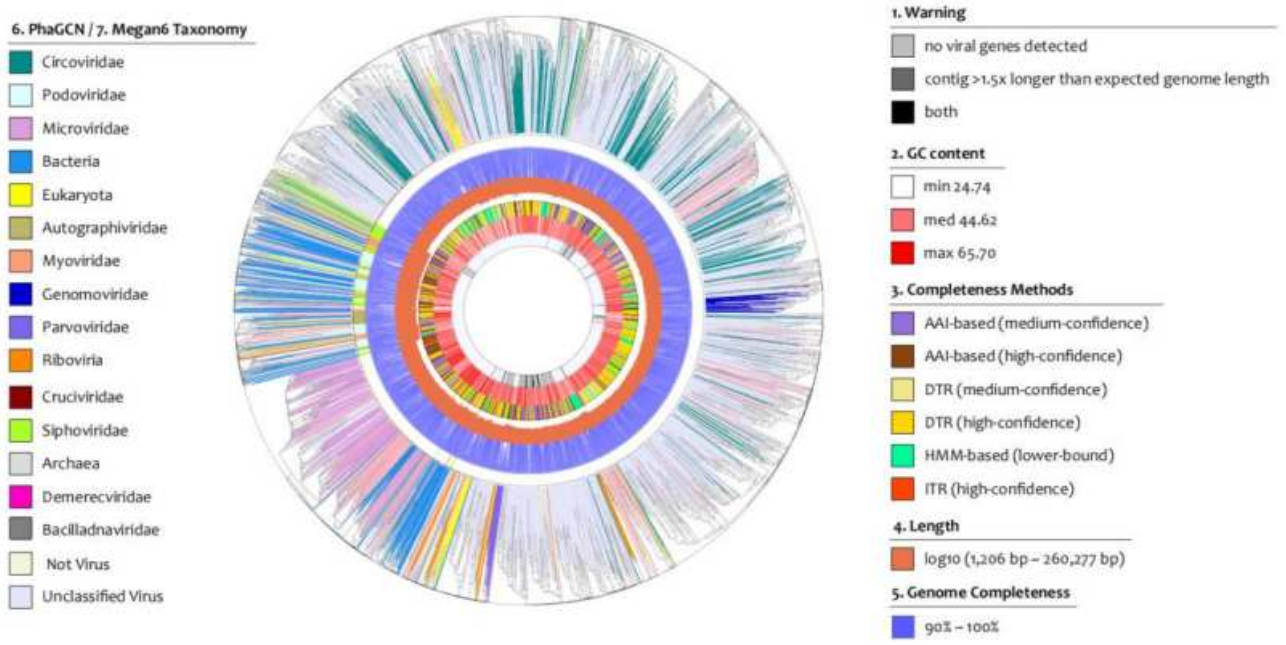

Figure 3

Viral proteomic phylogenetic tree of complete and near-complete viral genomes in the Dataset of Oyster Virome (DOV). The viral genomes were clustered based on their mutual amino acid identity using ViPTreeGen (v1.1.2). The layers from inside to outside show (1) the warning message of CheckV, (2) GC content of the viral genomes, (3) CheckV evaluation methods of genome completeness, (4) log10 value of genomic length, (5) percentage of genome completeness evaluated by CheckV, (6) viral families in order Caudovirales predicted by PhaGCN, and (7) viral families and non-viral annotations of all the genomes obtained by BLAST searches of the results from Diamond (v0.9.14.115) against the NCBI nonredundant protein sequence (nr) database (release November 2019).

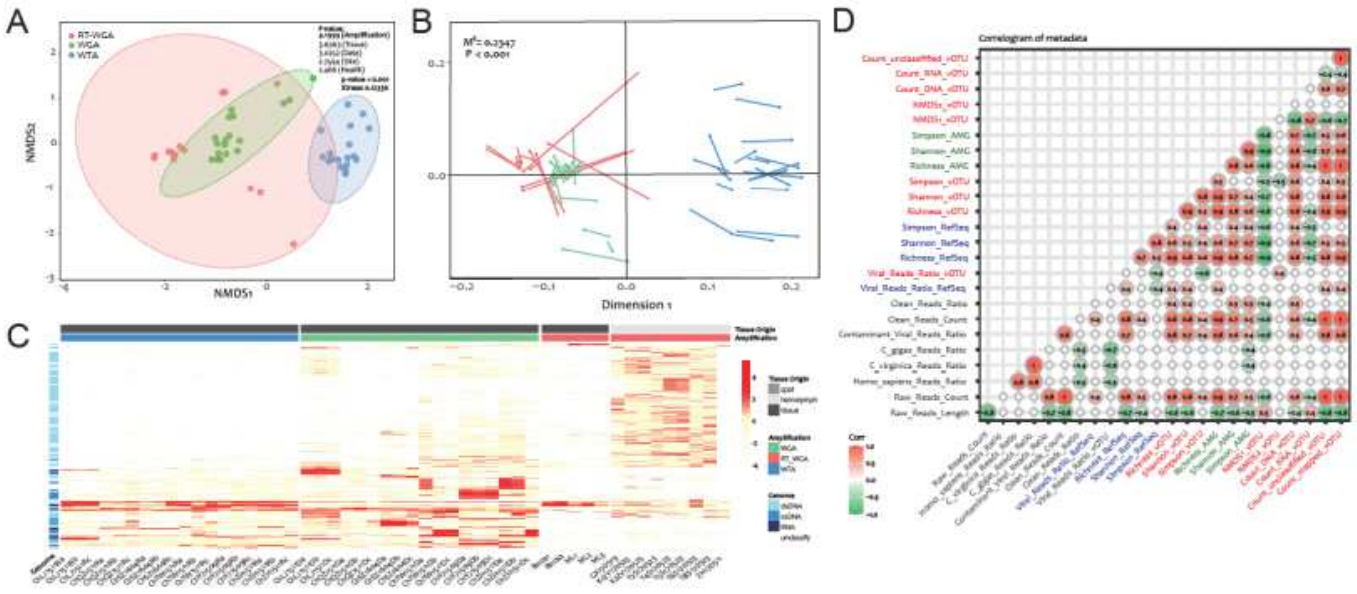

Figure 4

Viral community in the Dataset of Oyster Virome (DOV). (A) Nonmetric multidimensional scaling (NMDS) analysis shows the clusters of DOV libraries according to amplification groups. Nonparametric multivariate analysis of variance (permanova) was used. RT-WGA, reverse transcription and whole genome amplification; WGA, whole genome amplification; WTA whole transcriptome amplification. (B) Procrustes analysis of NMDS coordinates of viral communities based on comparisons of reference genomes (RefSeq, GOV, and IMG/VR) and de novo assembled viral contigs (vOTUs). (C) Heatmap of DOV vOTUs. The vOTUs clustered by the Euclidean method and colored by the viral genome types (dsDNA, ssDNA, RNA, and unclassified) are shown on the Y-axis. The DOV libraries ordered by amplification strategy (WGA, RT_WGA, and WTA) and tissue origin (hemolymph and mixed tissue) are shown on the X-axis. (D) Correlation matrix of oyster viral communities. Red labels (10), diversity indexes, viral reads ratio, and vOTU counts based on vOTUs mapping results; black labels (7), quality related parameters of library construction and sequencing; blue labels (4), diversity indexes and viral ratio based on the reference genomes (RefSeq, GOV, and IMG/VR) mapping results; green labels (3): diversity indexes based on the auxiliary metabolic genes (AMGs) mapping results.

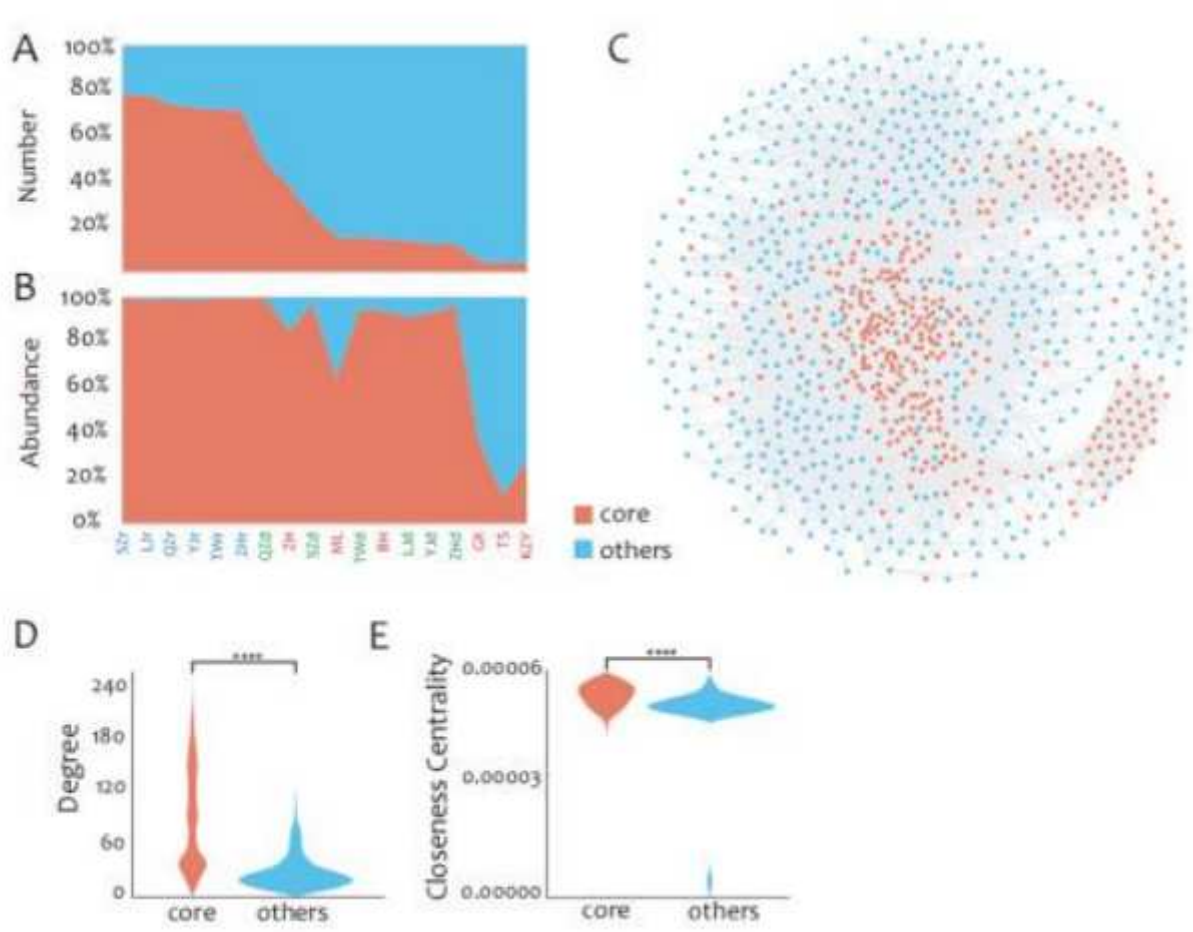

Figure 5

Core species (vOTUs) analysis of the Dataset of Oyster Virome (DOV). (A, B) Percentage of core vOTUs counts (A) and percentage of core vOTUs abundance (B) in different oyster viromic assembly groups (x-axis). Blue, whole transcriptome amplification (WTA) assembly groups; red, reverse transcription and whole genome amplification (RT-WGA) assembly groups; green, whole genome amplification (WGA) assembly groups. (C) Co-occurrence network of core (orange dots) and non-core (blue dots) vOTUs based on Spearman's correlation analysis (Spearman coefficient >0.7, p <0.01). (D, E) Topological parameters of nodes of degree (D) and closeness centrality (E)showing they were significantly different between the core and non-core vOTUs based on Wilcoxon rank sum test (**** P <0.0001)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile5.txt
- Additionalfile6.xlsx
- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- FigureS4.pdf
- FigureS5.pdf
- FigureS6.pdf
- FigureS7.pdf
- FigureS8.pdf
- FigureS9.pdf
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx