

***In silico* detection of a candidate disease-associated microorganism in metagenomic data from otter clams (*Lutraria rhynchaena* Reeve, 1854) affected by swollen siphon disease**

Phuong Thanh NGUYEN^{1,a}, Oanh Thi Phuong KIM^{1,2,b}, Binh Thi Nguyen LE^{1,c}, ✉

¹Vietnam Academy of Science and Technology, Institute of Biology, Nghia Do, Hanoi, Vietnam;

²Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Nghia Do, Hanoi, Vietnam;

^aORCID: 000-0002-1451-1317; ^bORCID: 0000-0003-3877-2375; ^cORCID: 0000-0002-7748-8628

ABSTRACT

Swollen siphon disease causes substantial mortality and economic losses in farmed snout otter clams (*Lutraria rhynchaena*) in Vietnam, yet its causative agent remains unknown. In this study, we performed metagenomic assembly and reconstructed metagenome-assembled genomes (MAGs) from available shotgun sequencing data of diseased siphon tissues. Assembly with metaSPAdes, viral identification with the geNomad pipeline, and genome binning with SemiBin yielded 219,737 contigs and four MAGs. Taxonomic profiling of assembled contigs revealed a bacteria-dominated metagenome (77%), primarily composed of *Pseudomonadota*, and an order-of-magnitude more annotations than previously reported. *Caudoviricetes* (~90%), a class of tailed bacteriophages, overwhelmingly dominated the viral contigs suggesting involvement of bacterial hosts associated with *Caudoviricetes* such as *Vibrio* spp., in disease etiology. Among the recovered MAGs, only one quality-sufficient MAG (bin 3; completeness 65.21%) was obtained and taxonomically assigned to the family *Simkaniaceae* (order *Chlamydiales*). Despite its partial genomic recovery, bin_3 encoded a near-complete type III secretion system and other intracellular lifestyle-associated features, including 247 putative virulence-associated genes. Our findings identify the *Simkaniaceae*-related bacterium as a candidate disease-associated microorganism. We hypothesize that it may compromise host immunity and facilitate opportunistic *Vibrio* infection, although this association is correlative and remains to be evaluated. Future experimental validation is needed to assess causality and inform disease prevention strategies.

Keywords: Metagenome, Molecular epidemiology, Pathogenicity, Snout otter clam, Aquaculture disease

✉Corresponding author: lenguyenbinh@ib.ac.vn

Received: 16.04.2026, Accepted: 26.08.2026, Publication Date: 01.10.2026

How to cite this article: Nguyen PT, Kim OTP, Le BTN (2026): *In silico* detection of a candidate disease-associated microorganism in metagenomic data from otter clams (*Lutraria rhynchaena* Reeve, 1854) affected by swollen siphon disease. Ankara Univ Vet Fak Derg, 73 (4), 468-481. DOI: 10.33988/auvfd.1927232.

Introduction

The snout otter clam (*Lutraria rhynchaena* Reeve, 1854) is an important bivalve species farmed widely in Vietnam, contributing to the development of local aquaculture markets (33, 46). In recent years, otter clam aquaculture in Vietnam has been severely disrupted due to an outbreak of swollen siphon disease, which is characterized by swelling inflammation, ulceration, and sloughing of siphon surface, leading to severe necrosis and death (16, 47). In addition to substantial economic losses to Vietnam's aquaculture

industry, the outbreak may pose environmental risks by affecting other natural bivalve species, and the causative pathogen remains elusive.

Diseases of cultured bivalves are often linked to complex microbial communities comprising bacteria, viruses, and obligate intracellular microorganisms. Bacterial taxa such as *Vibrio* spp. are frequently detected in diseased clams, although evidence for their role as primary pathogens remains inconsistent (17, 19, 45). Bivalve diseases may involve uncultured or poorly

characterized microorganisms that are not readily detected using conventional culture-based or targeted molecular diagnostics. For example, members of the order *Chlamydiales* have been widely reported in marine organisms and are commonly associated with tissue hypertrophy and epitheliocystis-like lesions (7, 35). In parallel, viral pathogens, particularly double-stranded DNA viruses of families *Herpesviridae* and *Iridoviridae*, are recognized as major drivers of disease outbreaks in bivalves aquaculture (3, 50). Regarding swollen siphon disease, various observational studies using electron microscopy have indicated the possible involvement of viral agents (16, 47). Metagenomic data from siphon tissues of infected snout otter clams have been first reported in 2020, primarily to characterize taxonomic abundance (28). Also, a novel circular single-stranded DNA (ssDNA) virus was subsequently detected in metagenomic data of diseased specimen (29). Despite these findings, a definitive causal agent has yet to be identified.

Advances in shotgun metagenomic sequencing and computational methods have enabled genome-resolved metagenomics, including the reconstruction of metagenome-assembled genomes (MAGs) (48). MAG-based analyses help direct characterization of dominant microbial taxa and their functional potential from host tissues, thereby improving the exploration of microbial diversity, functional metabolic repertoire, and host–microbe interactions (2, 23, 42). Applied to available metagenomic datasets, this approach holds promise for identifying key microbial taxa associated with pathological conditions.

In this study, we performed metagenomic assembly, annotation, and MAG characterization from an available metagenomic dataset of diseased snout otter clam (29). While the previous study focused on assembly-based identification of novel viral sequences, our present work aims to identify candidate disease-associated pathogens by characterizing MAG completeness, relative abundance, and metabolic functionality. By providing genomic insights into the microbiome of diseased clams, our study seeks to improve understanding of the microbial etiology of swollen siphon disease and support disease-prevention strategies for snout otter clam aquaculture.

Materials and Methods

Data acquisition and Preprocessing: Metagenomic sequencing data were retrieved from the NCBI Sequence Read Archive under accession SRP237889. According to the previous study, tissues from siphons of diseased snout otter clams (n=5) were dissected, pooled, homogenized and filtered through a 0.45 µm membrane, then treated with DNase to enrich viral particles by removing non–capsid-protected DNA. The processed material was used

for library preparation with the TruSeq Nano DNA Kit (Illumina) and sequenced on an Illumina HiSeq platform (29).

Our workflow employed RQCFilter2 from the BBTools suite to perform quality control of raw short-read data (10). Processing steps included quality trimming, removal of sequencing artifacts, linker and adapter trimming using BBDuk, followed by host and contaminant read removal using BBMap (10). High-quality reads were then assembled, structurally and functionally annotated, and binned into metagenome-assembled genomes (MAGs) using the standardized workflow detailed below.

Metagenomic Assembly: For metagenomic assembly of short-read dataset, interleaved paired-end FASTQ files were first error-corrected and reformatted using BBTools v39.03 (10), followed by *de novo* assembly with metaSPAdes v4.0.0 (38). Assembled contigs and scaffolds were standardized and summarized using AGP files conforming to the NCBI specification, and reads were mapped back to the final assemblies using bbmap to generate coverage information. These assembly outputs provided the contig-level foundation for downstream annotation and genome binning.

Structural and Functional Annotation: Assembled contigs were annotated following guidance from the JGI/IMG annotation pipeline (24). In brief, assembly was split into 10 Mb fragments for parallel processing, and each fragment underwent structural annotations using tRNAscan-SE, Infernal (Rfam), CRT, Prodigal, and GeneMarkS (4, 25), enabling comprehensive detection of protein-coding genes, non-coding sequences and other RNA structures.

Structural annotations from individual tools were merged into a consensus GFF file. Functional assignments were generated for this GFF file using LAST and HMMER against multiple protein family and domain databases, including SMART, COG, TIGRFAM, SUPERFAMILY, Pfam, and CATH FunFam (21, 27). Final outputs consisted of unified structural and functional annotation files in GFF format, along with tab-delimited summaries of gene products, KEGG Orthology terms, enzyme classifications, and scaffold-level lineage information.

For comprehensive exploration of viral and plasmid sequences, assembled metagenomic sequences were subjected to geNomad pipeline with default parameters (11), followed by CheckV v1.0.3 to assess genome completeness and confidence in the geNomad predictions (37). Taxonomic assignments are reported in accordance with ICTV guidelines.

MAG Reconstruction and Quality Assessment: Genome-resolved reconstruction was conducted using the metaMAGs workflow (v1.4.0), which integrates assembly outputs, read-mapping data, and functional annotations to recover MAGs. Contigs were clustered into genome bins using SemiBin v2.2.0 (41), and the resulting bins were subsequently refined using functional annotation files and, where available, contig-level lineage information. MAG quality was assessed using CheckM2 v1.0.2 (13), and bins were classified as high-quality, medium-quality, or low-quality according to MiMAG standards. Taxonomic classification of high- and medium-quality MAGs was performed using GTDB-Tk v2.2.0 (12).

Functional Profiling and Visualization: Recovered MAGs were functionally characterized using integrated KEGG Orthology annotations and metabolic module reconstruction. KO assignments were mapped to KEGG modules using ko_mapper.py from MicrobeAnnotator (43), and module completeness was calculated based on the presence of required functional steps. Visualization of metabolic potential and taxonomic composition across MAGs, via bar plots and Krona plots, were generated using KronaTools (40).

Comparative Genomic Screen for Virulence Determinants: To assess the pathogenic potential of the *Simkaniaceae*-related MAG, a reference catalogue of chlamydial virulence determinants (n=886) was compiled from the literature and public protein-family databases (Pfam, NCBIfam/TIGRFAM, COG, and KEGG Orthology). Candidate virulence genes were identified using complementary approaches, including annotation-based screening, profile-HMM searches with HMMER (27), secretion-system prediction with MacSyFinder (1), reciprocal orthology searches against representative chlamydial reference genomes using DIAMOND (9).

Results

Assembly of Metagenomic Data from Diseased Snout Otter Clam: From the processed reads, *de novo* assembly workflow using metaSPAdes v4.0.0 generated 219,737 contigs (which were later joined into 219,298 scaffolds), approximately 100,000 of which were at least 500 bp in length. The assembly achieved an N50 of 2,151 bp, and the longest contig reached 159,089 bp (Table 1). Notably, we recovered 3,934 contigs longer than 5 kb, of which 60 exceeded 50 kb. Given that the median genome sizes of ssDNA and dsDNA viruses are 3 kb and 40 kb, respectively, the assembly profile strongly suggests that our dataset may contain multiple near-complete viral genomes. However, further annotation and analyses were required to confirm the viral identities and their completeness.

Table 1. Summary of parameters from assembled metagenome

Parameters	Value
Number of contigs (≥ 0 bp)	219,737
Number of contigs (≥ 1000 bp)	37,577
Number of contigs (≥ 5000 bp)	3,934
Number of contigs (≥ 10000 bp)	1,334
Number of contigs (≥ 25000 bp)	383
Number of contigs (≥ 50000 bp)	60
Total length (≥ 0 bp)	196,788,588
Total length (≥ 1000 bp)	111,256,888
Total length (≥ 5000 bp)	47,043,184
Total length (≥ 10000 bp)	29,397,547
Total length (≥ 25000 bp)	15,621,811
Total length (≥ 50000 bp)	3,924,144
Number of contigs	98,749
Largest contig (bp)	159,089
Total length (bp)	154,812,446
GC (%)	45.82
N50	2,151
N90	666
Number of n per 100 kbp	0

Taxonomic assignment of assembled contigs delineated the metagenomic landscape of diseased snout otter clam tissues. Of the 85,528 contigs eligible for annotation, more than 84,000 were successfully assigned taxonomic information. Most annotated contigs were bacterial (77%), followed by viral (12%) and eukaryotic (10%) with archaeal contigs constituting a minor fraction (0.5%). Within the bacterial domain, the phylum *Pseudomonadota* predominated accounting for 47% of bacteria-assigned contigs, and encompassed three main classes: *Alphaproteobacteria*, *Betaproteobacteria*, and *Gammaproteobacteria*. Notably, the family *Vibrionaceae* (belonging to *Gammaproteobacteria*), which includes several *Vibrio* spp. implicated in bivalve diseases, accounted for approximately 1% of contigs assigned to bacteria. Additional abundant bacterial phyla included *Bacteroidota*, *Bacillota*, *Actinomycetota*, *Cyanobacteriota*, *Planctomycetota*, *Chlamydiota*, *Thermodesulfobacteriota*, *Chloroflexota*, and *Verrucomicrobiota*, each contributing between 1% and 17% of bacteria-assigned contigs (Figure 1). The viral component was dominated by *Caudoviricetes* (69% of viral-assigned contigs), a class of tailed bacteriophages, which was substantially more abundant than other classes, including *Malgrandaviricetes* (9%) and *Arfiviricetes* (8%). Meanwhile, most eukaryotic annotations were assigned to *Saccharomycetes* (93% of eukaryotic-assigned contigs), indicating probable yeast contamination introduced during sample handling, rather than representing endogenous clam-associated eukaryotes. Compared to the previous study, which reported taxonomic assignments for >8,000 contigs (29),

this analysis achieved nearly an order-of-magnitude increase in annotated sequences, yielding a far more comprehensive taxonomic landscape.

In-depth analysis of viral detection using geNomad pipeline (11), taxonomic evidences were successfully assigned to over 20,000 contigs, nearly 90% belonged to *Caudoviricetes* ($n = 17,905$) (Figure 2). As the virus enrichment protocol largely depleted the bacterial-sourced DNA, the abundance of this virus class indirectly hints at a diverse community of associated bacterial hosts in snout otter clams. Conversely, because the libraries were enriched for viral particles, they are expected to be biased against bacterial sequences. Hence the bacterial fraction recovered here, and any bacterial MAG reconstructed from it, represents only a partial view of the community and is incomplete. It is worth mentioning that, many lytic *Caudoviricetes* have been shown to target pathogenic *Vibrio* species, including *V. parahaemolyticus* (18). These observations imply that *Vibrio* infection may act as an additional contributor to the etiology of swollen siphon disease. This result may serve as a working hypothesis, as the present data are correlative and do not directly demonstrate *Vibrio*- or phage-host interactions.

Construction and Evaluation of Metagenome-Assembled Genomes (MAGs): The binning process for MAG construction yielded four bins exhibiting marked differences in assembly scale, fragmentation, and overall genome quality. Detailed characteristics of the resulting bins are presented in Table 2. Regarding assembly metrics, total assembly sizes span from ~0.16 Mb in the smallest bins (bin_1) to 4.35 Mb in the largest (bin_2), with corresponding numbers of contig ranging from 6 (bin_4) to 812 (bin_2). Indeed, bin_2 and bin_3 exhibit the largest genome sizes with the highest numbers of contigs and predicted features. The largest assembly (bin_2) comprises 812 sequences with median contig length of 4.6 kb, whereas smaller assembly such as bin_4 includes as few as six sequences but can consists of long contigs, reaching up to 159 kb and yielding a higher median contig length of 17.3 kb. Regarding gene prediction, the number of predicted features ranged from 218 in bin_1 to 2,740 in bin_2. Bin_2 and bin_3 contained the largest gene complements, with 2,740 and 988 genes, respectively. Notably, bin_3 exhibited the highest coding density, with predicted features occupying 80% of its assembled sequence (942,087 of 1,191,221 bp).

MAG quality was evaluated with CheckM2 using lineage-specific marker genes (13), providing a more direct measure of genome quality. Among the four bins, marker lineage assignment places one within Archaea (bin_1) and two within Bacteria (bin_2 and bin_3), while one bin (bin_4) remains unclassified beyond the root level, due to insufficient marker information for confident taxonomic assignment. Remarkably, completeness

determined by CheckM2 of bin_3 reached 65.21%, much higher than any other bins. This reflects the proportion of expected single-copy marker genes recovered and therefore approximately how much of a genome has been reconstructed. Nevertheless, a completeness of 65.21% indicates that one-third of the genome was not recovered; consequently, the metabolic reconstruction of this MAG is necessarily partial, and may reflect genome incompleteness rather than true absence. In the second place, bin_2 shows low completeness (23.7%) but higher contamination (7.45%), suggesting partial genome recovery. The completeness of bin_1 and bin_4 are 4% and 0%, respectively. Furthermore, bin_3 exhibited a low contamination level (0.68%) together with a high degree of strain heterogeneity (100%), making it the most informative bin recovered in our dataset. Consequently, based on MiMAG standards (8), only bin_3 was qualified as a medium-quality MAG, whereas the others were assigned as low-quality bins.

Functional Assessment and Taxonomic Composition of MAGs: The pathway annotation allows comparison of identified KEGG-ontology metabolic modules across all bins. Here, bin_2 and bin_3 show richer metabolic repertoire than bin_4, while no metabolic module is sufficiently detected in bin_1 (Figure 3). Modules belonging to the categories “Lipid metabolism,” “Fatty acid metabolism,” and “Central carbohydrate metabolism” were shared among the bins, with the greatest overlap observed between bin_2 and bin_3. Remarkably, bin_3 contained the highest number of modules with >80% completeness in the “Central carbohydrate metabolism” category. We noticed that bin_3 exhibits a diverse profile of metabolism-associated categories, with strong representation of carbohydrate metabolism and ATP synthesis compared to bin_2 (Figure 4). The co-occurrence of energy generation (ATP synthesis), core carbon processing, and nucleotide metabolism (represented by the “Pyrimidine metabolism” category) indicates a more comprehensive metabolic network. This observation is consistent with evaluation metrics indicating that bin_3 represents the most complete MAG in our dataset. However, because bin_3 is only ~65% complete, these module comparisons should be interpreted with caution.

Next, we sought to explore the taxonomic composition of contigs from the identified bins. Among the three low-quality bins, we only obtained meaningful results from bin_2, which was assigned completely to the genome of *Saccharomyces cerevisiae*, which probably represents a source of common contamination during handling of metagenomic samples (Figure 5). Meanwhile, as the sole medium-quality MAG, bin_3 contigs were overwhelmingly assigned to bacterial origin, particularly the class *Chlamydia*. Within this class, the family *Simkaniaceae* accounts for approximately 75% of total

annotations (n = 913), indicating that this lineage represents the core taxonomic signal (Figure 6). Smaller but notable fractions were attributed to *Parachlamydiaceae* (~6%) and other taxa (~4%), all of which belong to the class *Chlamydia*. This further reinforces the prominence of chlamydia-related taxa in bin_3 MAG. These results indicate that bin_3 represents a chlamydia-related bacterial genome, or a *Simkaniaceae*-like organism, which may correspond to a host-associated microbial lineage enriched in snout otter clams with swollen siphon disease. In the absence of a healthy-clam metagenome, this enrichment indicates disease association rather than absence or reduced abundance in healthy clams.

Detection of Protein Families in the *Simkaniaceae*-Related MAG: We focused our subsequent analyses on MAG bin_3 as the most promising candidate disease-associated microorganism. Contigs from bin_3 (n = 101) were aligned against three publicly available *Simkaniaceae* sequences obtained from NCBI, including two MAG isolates and the chromosome of *Simkania negevensis* Z. The results showed that, in every alignment case, all bin_3 contigs contained high-scoring segment regions that are mappable to *Simkaniaceae* reference sequences (Figure 7). Furthermore, phylogenetic assignment and reciprocal best-hit analyses consistently placed the organism as a sister lineage to *Simkania negevensis*, with genome-wide amino acid identities below the accepted genus boundary.

Notably, we identify that MAG bin_3 encoded an almost complete type III secretion system (T3SS), with 9

of 10 essential core components and 11 of 17 total components identified (Figure 8A). All detected components were supported by multiple independent lines of evidence and by orthology to *S. negevensis*, and were organized into three conserved operons, consistent with a functional secretion apparatus (Figure 8B). The few missing components (e.g., sctQ, sctK, sctW, sctI, sctF, sctP) are small, rapidly evolving proteins that are likely absent due to incomplete genome recovery.

The genome also encoded a complete F-like type IV secretion (conjugative) system comprising a single, contiguous 13.5 kb co-oriented *tra* operon (on contig 000380), a MOBF relaxase, and associated mobilization genes, a feature shared with *S. negevensis* but absent from several other representative *Chlamydiales* (Figure 8B, bottom strand). Genes were assigned to functional categories using a curated 886-entry *Chlamydiales* virulence-marker catalogue and scored by the number of mutually independent evidence layers supporting them (Pfam, TIGRFAM, COG and KEGG annotations; and reciprocal best-hit orthology to four chlamydial reference proteomes). We detected hallmark features of an obligate intracellular lifestyle (in category Host nutrient acquisition/energy parasitism), including five ATP:ADP antiporters, nutrient-acquisition genes, 45 proteins with eukaryotic-like domains, and six high-confidence inclusion membrane (Inc) proteins. In total, 247 genes of bin_3 showed evidence of virulence-related functions, including 60 high-confidence candidates, which are all summarized in Table 3.

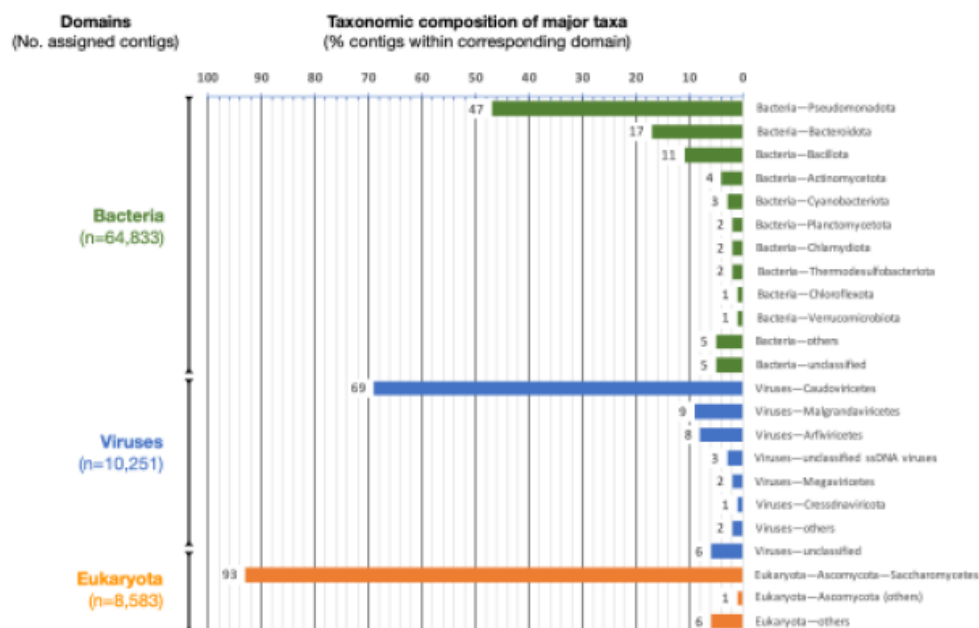


Figure 1. Overview of taxonomic composition of annotated contigs (n=85,528). Taxa accounting for <1% within each domain are merged into the “Others” category

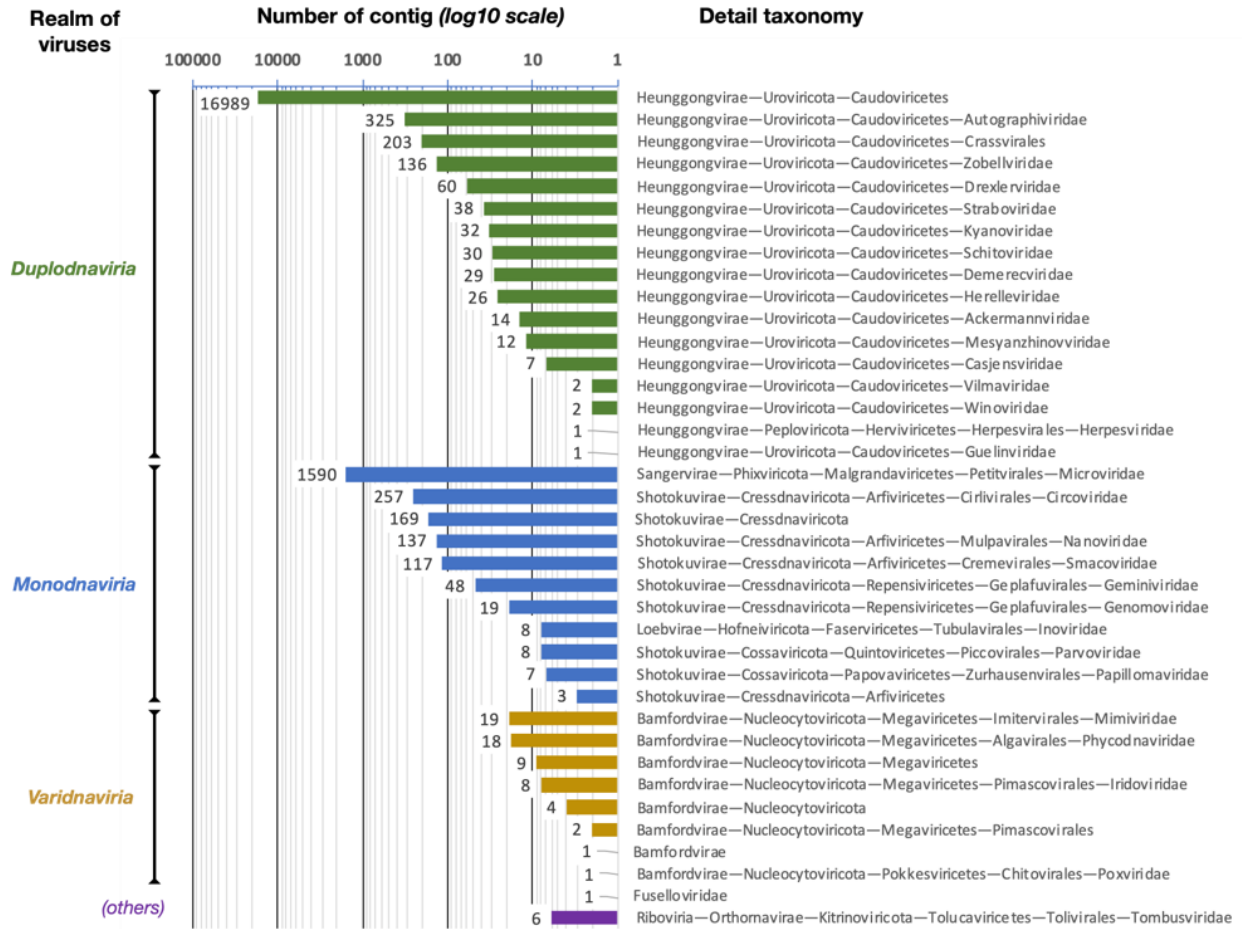


Figure 2. Summary of the number of contigs by viral taxonomic classification. Viruses are organized into main realms (i.e., Duplodnaviria, Monodnaviria, Varidnaviria and Riboviria)

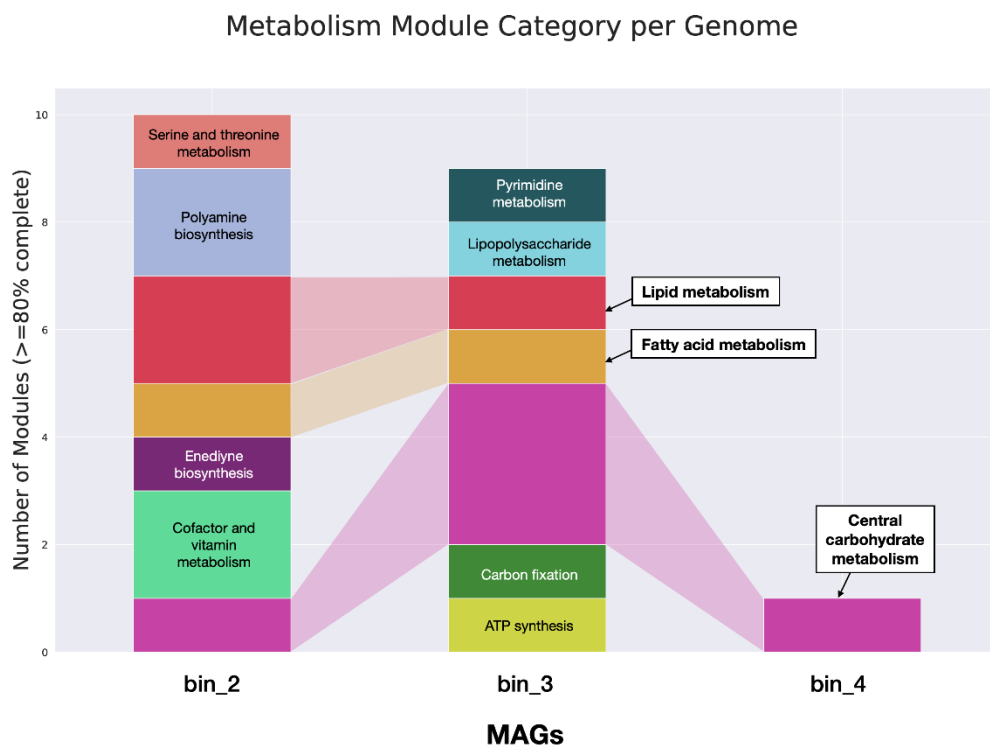


Figure 3. Metabolic module categories (restricted to modules with >80% completeness) recovered from the identified bins. Categories shared between bins are linked by lighter shading.

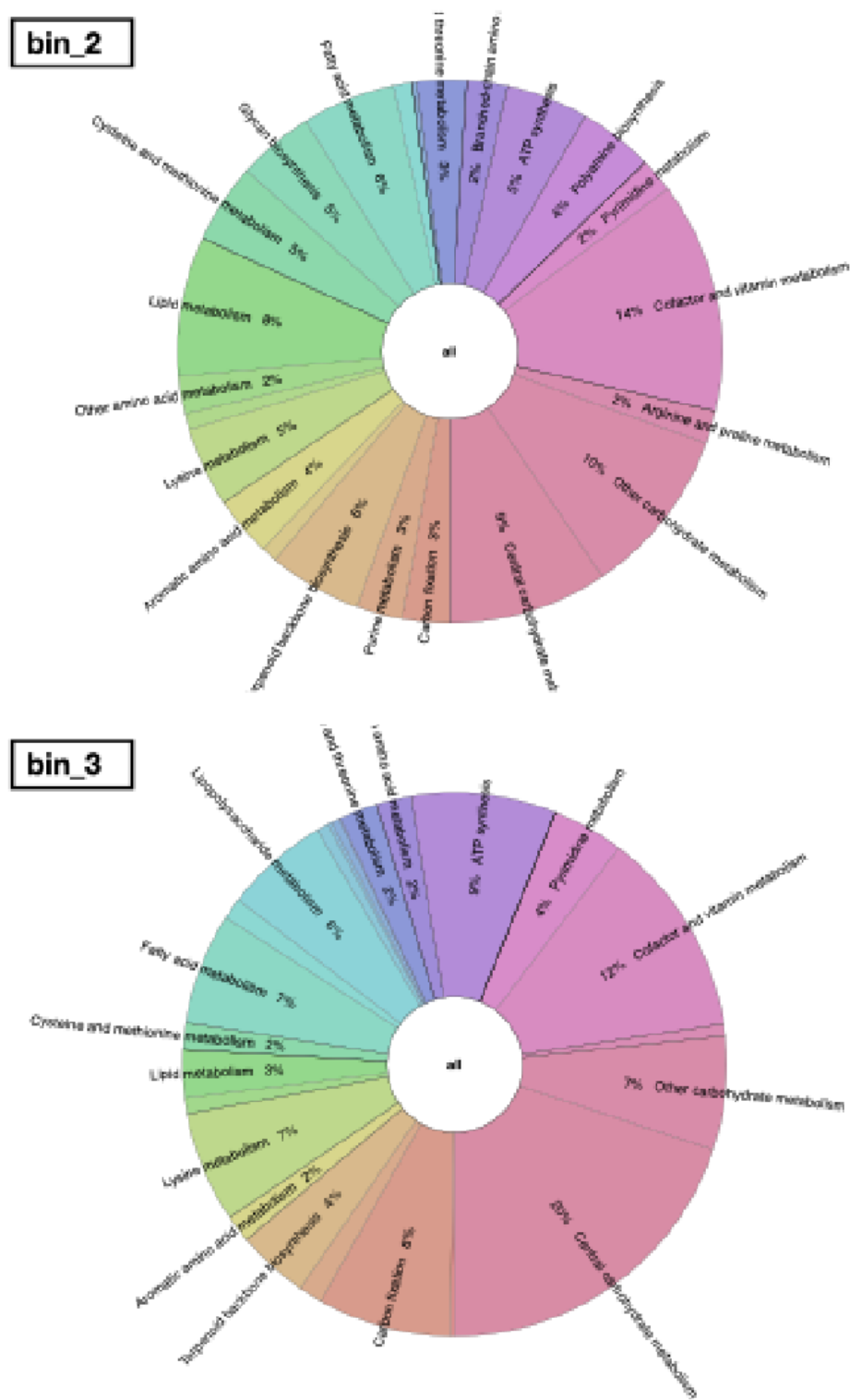


Figure 4. Composition of major metabolic modules identified from bin_2 and bin_3

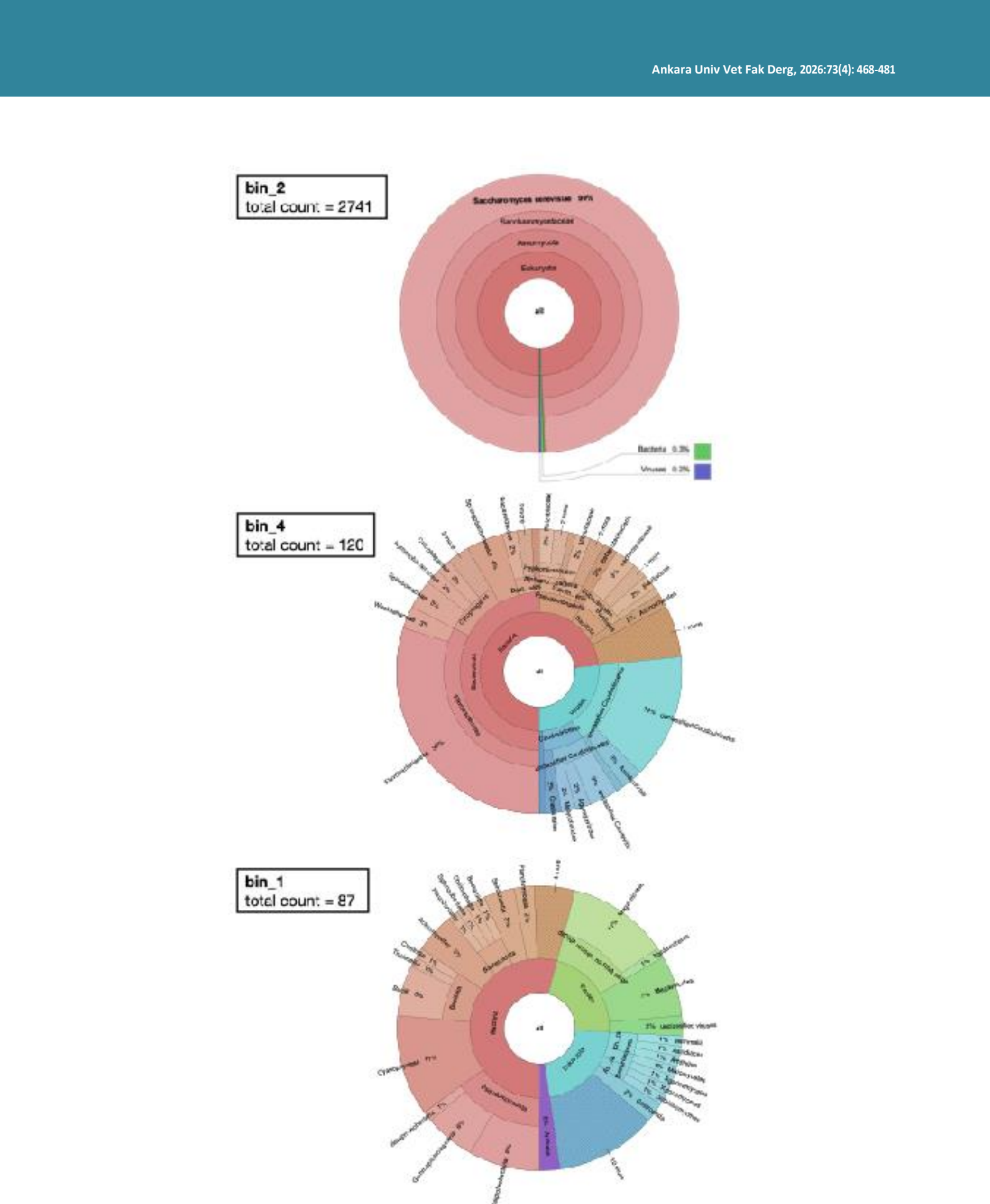


Figure 5. Krona plots showing taxonomic compositions of low-quality bins (from top to bottom: bin_2, bin_4 and bin_1)

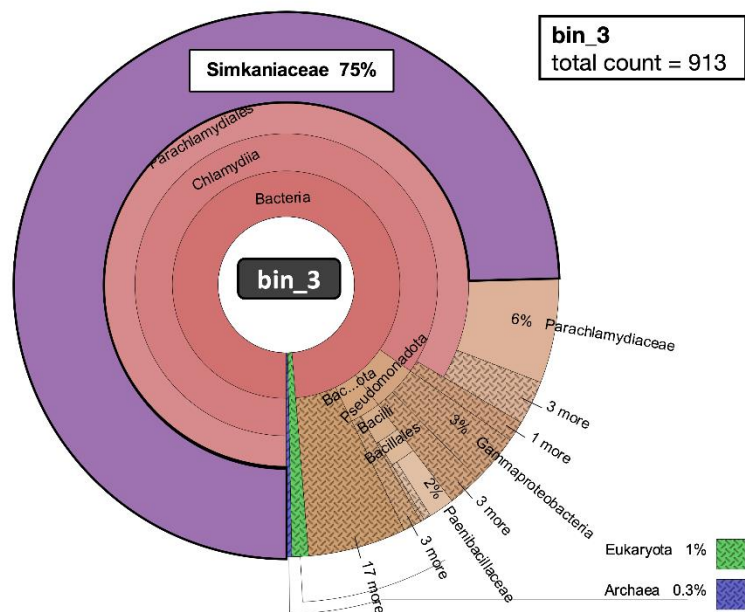
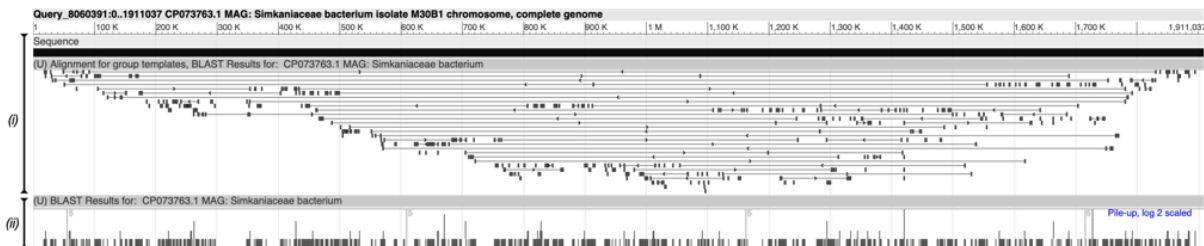
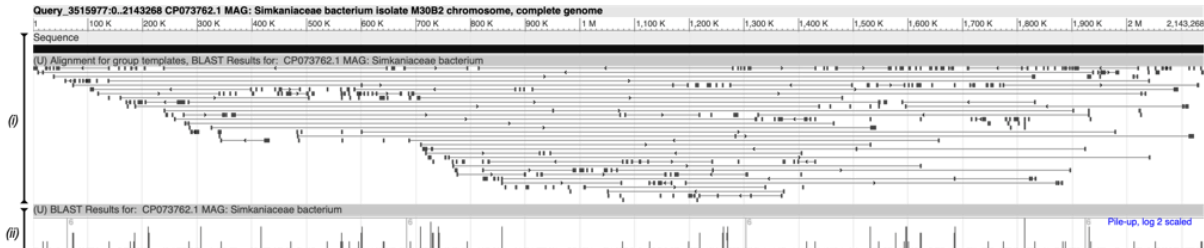


Figure 6. Summary of taxonomic assignments from contigs of bin_3 MAG. The family *Simkaniaceae* (class *Chlamydia*) represents the dominant assignment (~75%; n = 913), supporting the identification of bin_3 as a *Simkaniaceae*-like genome.

CP073763.1 MAG: *Simkaniaceae* bacterium isolate M30B1 chromosome (1.911 Mb)



CP073762.1 MAG: *Simkaniaceae* bacterium isolate M30B2 chromosome (2.143 Mb)



NC_015713.1 *Simkania negevensis* Z chromosome (2.496 Mb)

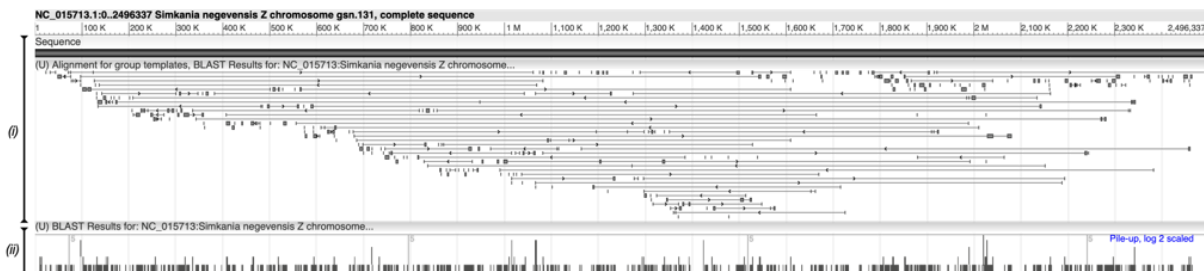


Figure 7. Pile-up visualization of BLASTn alignments of bin_3 contigs (n=101) against three NCBI *Simkaniaceae*-related reference sequences. (i) Individual alignment of each contig showing high-score pair regions; (ii) Cumulative coverage over reference sequences from all alignments.

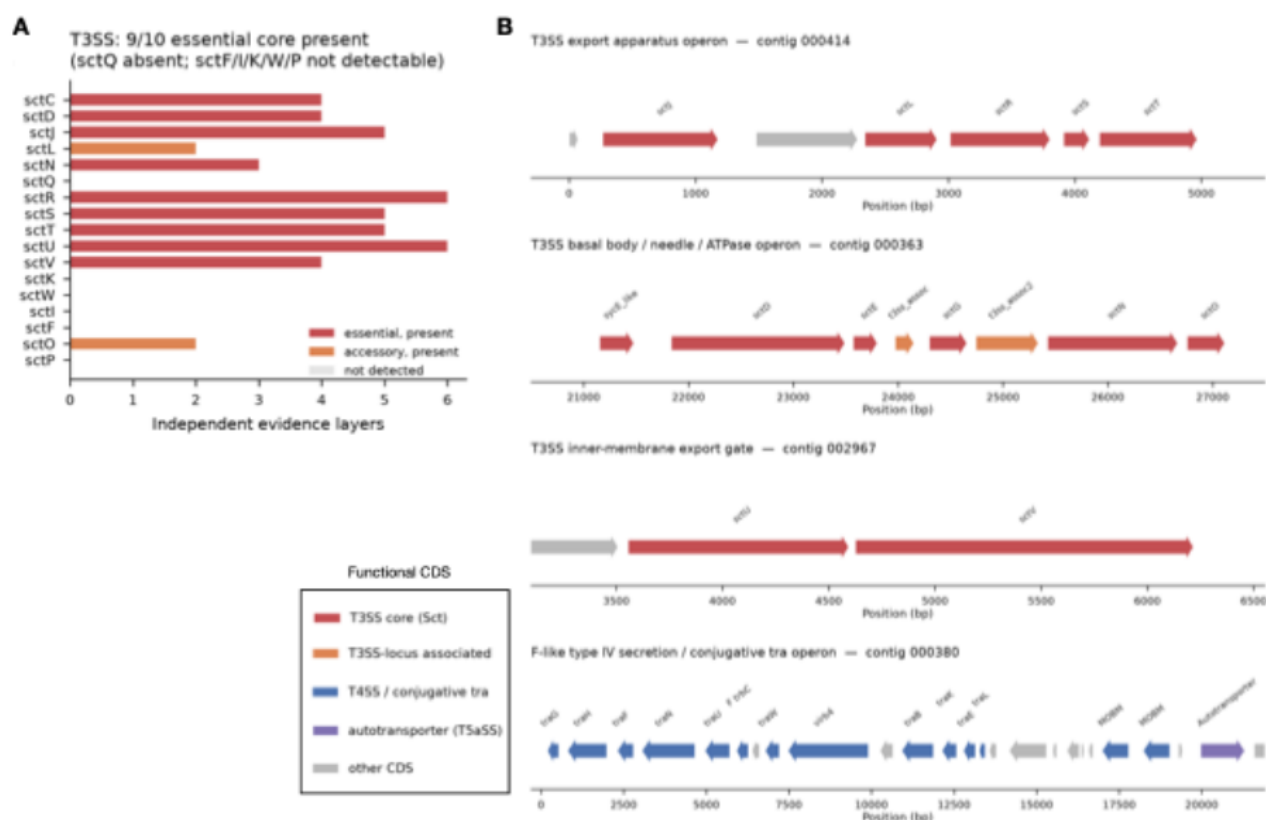


Figure 8. Detection of the type III and F-like type IV secretion systems in the Simkaniaceae MAG (bin_3). (A) Completeness of the type III secretion system (T3SS). Bars indicate the number of independent lines of evidence supporting each T3SS component based on genome annotation, profile-HMM searches, secretion-system prediction, and orthology. (B) Genomic organization of the secretion-system loci, containing three T3SS loci and one F-like T4SS locus, each organized as contiguous, co-oriented gene clusters. Arrows represent predicted coding sequences (scaled), genes (functional CDS) are colored as described in the legend box.

Table 2. Statistics of identified bins from metagenomic data

Bin IDs		bin_1	bin_2	bin_3	bin_4
Processed sequences statistics	Number of seqs	18	812	101	6
	Total length (bp)	162,975	4,351,129	1,191,221	240,416
	Length range (shortest-longest) (bp)	3,394 – 28,152	3,000 – 20,552	3,281 – 39,154	5,097 – 159,089
	Median length (bp)	7,361	4,592	9,995	17,324.5
Predicted genes statistics (Prodigal v2.6.3 CDS prediction)	Number of predicted features	218	2,740	988	353
	Number of seqs	18	801	101	6
	Total length (bp)	151,989	3,127,230	942,087	218,100
	Length range (shortest-longest) (bp)	123 – 3,513	75 – 7,887	90 – 5,046	84 – 6,906
	Median length (bp)	585	897	822	375
MAGs evaluation (CheckM2 assessments)	Marker lineage	<i>k_Archaea</i> (UID2)	<i>k_Bacteria</i> (UID203)	<i>k_Bacteria</i> (UID2982)	<i>root</i> (UID1)
	# genomes	207	5,449	88	5,656
	# markers	149	104	230	56
	# marker sets	107	58	148	24
	Completeness (%)	4.03	23.68	65.21	0
	Contamination (%)	0	7.45	0.68	0
	Strain heterogeneity (%)	0	3.45	100	0

Table 3. Summary of virulence-determinant screening of the *Simkaniaceae* MAG (bin_3)

Functional category	Genes	Virulence-marker evidence (*) (High/Med/Low)	Median evidence layers	<i>S. negevensis</i> orthologues (median AAI)
Type III secretion (T3SS)	34	16 / 10 / 8	2	31 (64.2%)
Type IV secretion / conjugation	30	24 / 4 / 2	2	25 (49.9%)
Type II secretion / type IV pilus / Tad	24	9 / 6 / 9	2	22 (53.6%)
Type V secretion (autotransporters, Pmp-like adhesins)	9	1 / 4 / 4	3	4 (34.6%)
Sec / Tat general export	21	9 / 8 / 4	2	17 (61.3%)
Inclusion membrane (Inc) candidates	5	0 / 3 / 2	1	0 (0%)
Eukaryotic-like domains (candidate effectors)	45	8 / 20 / 17	2	26 (50.1%)
Host nutrient acquisition / energy parasitism	35	5 / 17 / 13	2	27 (50.5%)
Iron acquisition	6	0 / 3 / 3	1	6 (52.3%)
Toxins / membrane lysis	6	1 / 2 / 3	1	4 (44.4%)
Virulence-associated proteases	11	1 / 5 / 5	2	10 (53.6%)
Envelope / LPS / peptidoglycan	13	1 / 9 / 3	2	13 (48.3%)
Stress response / chaperones	10	0 / 8 / 2	2	9 (69.1%)
Regulatory	18	4 / 5 / 9	1	14 (69.6%)

(*) Confidence tiers: high, an annotated secretion-system component or three concordant layers with orthology assigned; medium, two or more concordant layers; low, a single layer.

Discussion and Conclusion

In this study, the comprehensive analysis of metagenomic data derived from snout otter clams affected by swollen siphon disease provided deep insights into the diversity and enabled genomic characterization of candidate disease-associated microorganisms. Notably, via MAG reconstruction, we recovered a single medium-quality bacterial MAG belonging to the family *Simkaniaceae* (order *Chlamydiales*), which accounted for approximately 75% of taxonomic signal within the respective MAG (bin_3). This observation suggests an association with the disease state and is consistent with patterns reported for uncultured chlamydial pathogens in studies of similar disease etiology (30, 35, 49).

It is noteworthy that the employment of the updated *in silico* workflow in our study produced a ten-fold increase of annotated sequences compared to the previous analysis (29), providing a markedly more comprehensive taxonomic characterization. Our virus-focused analysis of the assembled metagenome identified *Caudoviricetes* as the dominant viral taxon, accounting for 90% of virus-assigned contigs. This observation implicates *Vibrio* spp., well-recognized hosts of *Caudoviricetes* phages, as potential contributors to swollen siphon disease (18). However, previous experimental challenge studies using *Vibrio* spp. alone failed to reproduce the swollen siphon disease phenotype in clams (47), implying that *Vibrio* infection may play a secondary role. Regarding the detection of a *Simkaniaceae* MAG, we hypothesize that a primary intracellular infection by this organism may compromise host tissue integrity or immune function, allow opportunistic *Vibrio* proliferation, and resulting in a

bloom of *Caudoviricetes* phages. This proposed *Simkaniaceae*–*Vibrio*–*Caudoviricetes* interaction scheme should be regarded as a hypothesis, as it is based solely on co-occurrence patterns from a single pooled sample and lacks direct quantitative or statistical support. Further investigations will be required to evaluate this disease model.

In marine systems, *Simkaniaceae* are obligate intracellular bacteria family with broad pathogenic potential, including bivalves, teleost fish, and corals (22). In bivalves, *Simkaniaceae*-like organisms are linked to tissue hypertrophy syndromes, exemplified by the uncultured strain “Maxima-CLO” associated with oedema disease in pearl oyster (*Pinctada maxima*) (15, 26). Related to *Simkaniaceae*, many members of the *Chlamydiia* class, such as *Parachlamydia* and *Rhabdochlamydia*, or *Chlamydia trachomatis*, are well-known animal pathogens, especially in human respiratory diseases (31, 32). Moreover, some chlamydial lineages infect microbial eukaryotes and possess more extensive metabolic capabilities (14).

In the current study, *Simkaniaceae* MAG (bin_3) similarly dominates diseased siphon tissue and possesses metabolic pathways for ATP synthesis, central carbohydrate metabolism, carbon fixation, lipid/fatty acid metabolism, pyrimidine metabolism, and lipopolysaccharide metabolism, which is consistent with an active metabolic network. Nonetheless, because one-third of the genome was not recovered, this metabolic profile is still incomplete, and modules that appear absent cannot be interpreted as genuine metabolic gaps. Compared with classical members of the family *Chlamydiaceae*, which

are highly dependent on host-derived energy, many environmental and marine *chlamydiae* exhibit expanded metabolic capabilities. Genomic analyses of *Simkaniaceae* and related *Chlamydia*-like organisms have demonstrated the presence of complete or near-complete TCA cycles, glycolytic pathways, and components of oxidative phosphorylation, indicating a greater degree of metabolic autonomy (14, 49). This metabolic versatility facilitates persistence across diverse host taxa and environmental conditions. Consistently, the co-occurrence of pathways such as ATP synthesis, central carbohydrate metabolism, lipid, and fatty-acid metabolism in our results suggests the hypothesis that this organism is not merely surviving within host tissue but actively modifying the intracellular environment. Such metabolic activity may contribute to the tissue hypertrophy and swelling characteristics of swollen siphon disease. Importantly, these functional insights are reported here for the first time and were not included in the previous analyses (29).

Beyond bivalves, *Simkaniaceae*-like organisms are well-established agents associated with epitheliocystis in marine fish, including *Candidatus Syngnamydia salmonis* in Atlantic salmon (39), and *Candidatus Syngnamydia venezia* in broad-nosed pipefish (20). They are recognized as emerging transboundary threats in fish aquaculture (6, 44). In reef-building corals such as *Pocillopora acuta*, *Simkaniaceae* occur within cell-associated microbial aggregates and may proliferate under environmental stress, contributing to tissue loss syndromes (34). Collectively, these evidence supports *Simkaniaceae* as adaptable marine opportunists with pathogenic potential, underscoring the need for targeted diagnostics and surveillance in aquaculture.

Members of the *Chlamydiales*, including *Simkania negevensis*, typically rely on a type III secretion system (T3SS) together with an array of secreted effectors and inclusion-membrane proteins to invade host cells, establish an intracellular niche, and manipulate host signaling (14, 49). We identified the near-complete T3SS, predicted Inc proteins, ATP:ADP antiporters, and eukaryotic-like proteins in the *Simkaniaceae* MAG, these findings are consistent with genomic adaptations associated with the intracellular lifestyle of *Chlamydiales*. In detail, chlamydial T3SSs deliver effector proteins into host cells, while Inc proteins modify the inclusion membrane and mediate interactions with host-cell pathways required for intracellular survival (5, 36). The complete F-like conjugative system is also notable because *Simkania negevensis* carries a related conjugative element, suggesting that horizontal gene transfer may contribute to the acquisition and diversification of accessory functions in this lineage (14). Together, these features are consistent with the potential for intracellular host interactions, but do not necessarily imply

pathogenicity. Future studies should combine high-coverage, non-enriched metagenomic sequencing with quantitative and experimental validations (such as qPCR, *in situ* localization and controlled infection assays) to confirm the reconstructed MAGs, determine whether the suspected lineage is associated with disease, and evaluate its pathogenic potential.

In conclusion, we successfully assembled and analyzed metagenomic sequences derived from swollen siphon affected snout otter clams, providing genome-resolved insights into the microbial landscape associated with the disease state. Viral profiling revealed the pronounced dominance of *Caudoviricetes*, indicating that bacteriophages may play a role in disease-associated microbial dynamics. Importantly, we reconstructed and characterized a novel *Simkaniaceae*-related MAG. Despite its incomplete recovery, the genome retained multiple hallmarks of an intracellular lifestyle, including a near-complete type III secretion system, which supports its potential for host interaction. The metabolic potential encoded by this MAG suggests that a *Simkaniaceae*-like intracellular organism may be associated with disease-related disruptions of host cellular homeostasis. Our study demonstrates the power of genome-resolved metagenomics to uncover previously uncharacterized microorganisms with potential relevance to aquaculture studies.

Financial Support

This work was funded by the Vietnam National Foundation for Science and Technology Development under grant number 106.05-2020.20.

Ethical Statement

This *in silico* study does not involve any direct animal experimentation.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

BTNL, OTPK and PTN conceived the idea and planned the manuscript. BTNL contributed to sample preparation. PTN and OTPK have provided significant scientific support and contributed to the interpretation of the results. All authors contributed significantly, providing feedback and helping to shape the manuscript.

Data Availability Statement

Raw metagenomic sequencing data was retrieved from the NCBI Sequence Read Archive under accession SRP237889. The analysis data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative Artificial Intelligence

The authors declare that generative artificial intelligence (AI) and AI-assisted technologies were not used in the writing of this manuscript or in the preparation of figures.

References

1. Abby SS, Néron B, Ménager H, et al (2014): *MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems*. PLOS ONE, **9**, e110726. DOI: 10.1371/journal.pone.0110726.
2. Banerjee G, Papri SR, Banerjee P (2024): *Protocol for the construction and functional profiling of metagenome-assembled genomes for microbiome analyses*. STAR Protoc, **5**, 103167. DOI: 10.1016/j.xpro.2024.103167.
3. Batista FM, Arzul I, Pepin J-F, et al (2007): *Detection of ostreid herpesvirus 1 DNA by PCR in bivalve molluscs: a critical review*. J Virol Methods, **139**, 1–11. DOI: 10.1016/j.jviromet.2006.09.005.
4. Besemer J, Lomsadze A, Borodovsky M (2001): *GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions*. Nucleic Acids Res, **29**, 2607–2618. DOI: 10.1093/nar/29.12.2607.
5. Betts-Hampikian HJ, Fields KA (2010): *The chlamydial type iii secretion mechanism: Revealing cracks in a tough nut*. Front Microbiol, **1**, 2010. <https://doi.org/10.3389/fmicb.2010.00114> DOI: 10.3389/fmicb.2010.00114.
6. Blandford MI, Taylor-Brown A, Schlacher TA, et al (2018): *Epitheliocystis in fish: An emerging aquaculture disease with a global impact*. Transbound Emerg Dis, **65**, 1436–1446. DOI: 10.1111/tbed.12908.
7. Bower SM, McGladdery SE, Price IM (1994): *Synopsis of infectious diseases and parasites of commercially exploited shellfish*. Annual Review of Fish Diseases, **4**, 1–199. DOI: 10.1016/0959-8030(94)90028-0.
8. Bowers RM, Kyrpidis NC, Stepanauskas R, et al (2017): *Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea*. Nat Biotechnol, **35**, 725–731. DOI: 10.1038/nbt.3893.
9. Buchfink B, Xie C, Huson DH (2015): *Fast and sensitive protein alignment using DIAMOND*. Nat Methods, **12**, 59–60. DOI: 10.1038/nmeth.3176.
10. Bushnell B (2014): *BBMap: A Fast, Accurate, Splice-Aware Aligner*. LBL Publications.
11. Camargo AP, Roux S, Schulz F, et al (2024): *Identification of mobile genetic elements with geNomad*. Nat Biotechnol, **42**, 1303–1312. DOI: 10.1038/s41587-023-01953-y.
12. Chaumeil P-A, Mussig AJ, Hugenholtz P, et al (2020): *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database*. Bioinformatics, **36**, 1925–1927. DOI: 10.1093/bioinformatics/btz848.
13. Chklovski A, Parks DH, Woodcroft BJ, et al (2023): *CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning*. Nat Methods, **20**, 1203–1212. DOI: 10.1038/s41592-023-01940-w.
14. Collingro A, Tischler P, Weinmaier T, et al (2011): *Unity in variety--the pan-genome of the Chlamydiae*. Mol Biol Evol, **28**, 3253–3270. DOI: 10.1093/molbev/msr161.
15. Croxford M, Jones B (2012): *Aquatic Animal Health Subprogram: Investigation of Chlamydiales-like organisms in pearl oysters, Pinctada maxima FRDC Project 2008/031*. Government of Western Australia Department of Fisheries, Perth. Report Fisheries Research Report No. 232.
16. Dang T, Phan T, Nguyen T (2018): *Detection of virus-like particles (VLPs) in otter clam (Lutraria philippinarum Reeve, 1854) spat collected from hatcheries*. Vietnam J Agric Sci, **16**, 867–873.
17. Destoumieux-Garzón D, Canesi L, Oyanedel D, et al (2020): *Vibrio-bivalve interactions in health and disease*. Environ Microbiol, **22**, 4323–4341. DOI: 10.1111/1462-2920.15055.
18. Duarte J, Trindade D, Oliveira V, et al (2024): *Isolation and Characterization of Infection of Four New Bacteriophages Infecting a Vibrio parahaemolyticus Strain*. Antibiotics, **13**, 1086. DOI: 10.3390/antibiotics13111086.
19. Dubert J, Barja JL, Romalde JL (2017): *New insights into pathogenic vibrios affecting bivalves in hatcheries: Present and future prospects*. Front Microbiol, **8**, 762. DOI: 10.3389/fmicb.2017.00762.
20. Fehr A, Walther E, Schmidt-Posthaus H, et al (2013): *Candidatus Syngnamydia venezia, a novel member of the phylum chlamydiae from the broad nosed pipefish, syngnathus typhle*. PLoS One, **8**, e70853. DOI: 10.1371/journal.pone.0070853.
21. Frith MC, Wan R, Horton P (2010): *Incorporating sequence quality data into alignment improves DNA read mapping*. Nucleic Acids Research, **38**, e100–e100. DOI: 10.1093/nar/gkq010.
22. Greub G, Raoult D (2002): *Parachlamydiaceae: potential emerging pathogens*. Emerg Infect Dis, **8**, 625–630. DOI: 10.3201/eid0806.010210.
23. Handelsman J (2004): *Metagenomics: application of genomics to uncultured microorganisms*. Microbiol Mol Biol Rev, **68**, 669–685. DOI: 10.1128/MMBR.68.4.669-685.2004.
24. Huntemann M, Ivanova N (2023): *IMG Annotation Pipeline (IMGAP) v5.1.13*. DOI: 10.11578/DC.20240716.3.
25. Hyatt D, Chen G-L, Locascio PF, et al (2010): *Prodigal: prokaryotic gene recognition and translation initiation site identification*. BMC Bioinformatics, **11**, 119. DOI: 10.1186/1471-2105-11-119.
26. Itoiz S, Mouronvalle C, Perennou M, et al (2023): *Co-infection of two eukaryotic pathogens within clam populations in Arcachon Bay*. Front Microbiol, **14**, 1250947. DOI: 10.3389/fmicb.2023.1250947.
27. Johnson LS, Eddy SR, Portugaly E (2010): *Hidden Markov model speed heuristic and iterative HMM search procedure*. BMC Bioinformatics, **11**, 431. DOI: 10.1186/1471-2105-11-431.
28. Kagaya Y, Minei R, Duong HTT, et al (2020): *Metagenome Sequences from the environment of diseased otter clams, Lutraria rhynchaena, from a farm in Vietnam*. Microbiol Resour Announc, **9**, e01068-19. DOI: 10.1128/MRA.01068-19.

29. Kim OTP, Kagaya Y, Tran HS, et al (2020): A novel circular ssDNA virus of the phylum Cressdnaviricota discovered in metagenomic data from otter clams (*Lutraria rhynchaena*). Arch Virol, **165**, 2921–2926. DOI: 10.1007/s00705-020-04819-9.
30. Lamothe F, Greub G (2010): Amoebal pathogens as emerging causal agents of pneumonia. FEMS Microbiol Rev, **34**, 260–280. DOI: 10.1111/j.1574-6976.2009.00207.x.
31. Lamothe F, Jaton K, Vaudaux B, et al (2011): Parachlamydia and Rhabdochlamydia: emerging agents of community-acquired respiratory infections in children. Clin Infect Dis, **53**, 500–501. DOI: 10.1093/cid/cir420.
32. Lieberman D, Kahane S, Lieberman D, et al (1997): Pneumonia with serological evidence of acute infection with the Chlamydia-like microorganism “Z.” Am J Respir Crit Care Med, **156**, 578–582. DOI: 10.1164/ajrccm.156.2.9608081.
33. Luca M, Nam DX (2012): Hatchery techniques applied for the artificial production of snout otter clam (*Lutraria rhynchaena*) in small scale farms in Nha Trang City, Vietnam. Aquac Asia, **17**, 25–29.
34. Maire J, Tandon K, Collingro A, et al (2023): Colocalization and potential interactions of Endozoicomonas and chlamydiae in microbial aggregates of the coral Pocillopora acuta. Sci Adv, **9**, eadg0773. DOI: 10.1126/sciadv.adg0773.
35. Meijer A, Roholl PJM, Ossewaarde JM, et al (2006): Molecular evidence for association of chlamydiales bacteria with epitheliocystis in leafy seadragon (*Phycodurus eques*), Silver Perch (*Bidyanus bidyanus*), and Barramundi (*Lates calcarifer*). Appl Environ Microbiol, **72**, 284–290. DOI: 10.1128/AEM.72.1.284-290.2006.
36. Mital J, Miller NJ, Dorward DW, et al (2013): Role for chlamydial inclusion membrane proteins in inclusion membrane structure and biogenesis. PLoS One, **8**, e63426. DOI: 10.1371/journal.pone.0063426.
37. Nayfach S, Camargo AP, Schulz F, et al (2021): CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol, **39**, 578–585. DOI: 10.1038/s41587-020-00774-7.
38. Nurk S, Meleshko D, Korobeynikov A, et al (2017): metaSPAdes: a new versatile metagenomic assembler. Genome Res, **27**, 824–834. DOI: 10.1101/gr.213959.116.
39. Nylund S, Steigen A, Karlsbakk E, et al (2015): Characterization of “Candidatus Syngnamydia salmonis” (Chlamydiales, Simkaniaceae), a bacterium associated with epitheliocystis in Atlantic salmon (*Salmo salar* L.). Arch Microbiol, **197**, 17–25. DOI: 10.1007/s00203-014-1038-3.
40. Ondov BD, Bergman NH, Phillippy AM (2011): Interactive metagenomic visualization in a Web browser. BMC Bioinformatics, **12**, 385. DOI: 10.1186/1471-2105-12-385.
41. Pan S, Zhao X-M, Coelho LP (2023): SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. Bioinformatics, **39**, i21–i29. DOI: 10.1093/bioinformatics/btad209.
42. Quince C, Walker AW, Simpson JT, et al (2017): Shotgun metagenomics, from sampling to analysis. Nat Biotechnol, **35**, 833–844. DOI: 10.1038/nbt.3935.
43. Ruiz-Perez CA, Conrad RE, Konstantinidis KT (2021): MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinformatics, **22**, 11. DOI: 10.1186/s12859-020-03940-5.
44. Stride MC, Polkinghorne A, Nowak BF (2014): Chlamydial infections of fish: diverse pathogens and emerging causes of disease in aquaculture species. Vet Microbiol, **170**, 19–27. DOI: 10.1016/j.vetmic.2014.01.022.
45. Thuy Pham T, Hong Nhi Ho T, Nguyen VD (2014): Screening for bacteriocin-like antimicrobial activity against shrimp pathogenic vibrios and molecular identification of marine bacteria from otter clam *lutraria philippinarum*. Thai J Vet Med, **44**, 345–353. DOI: 10.56808/2985-1130.2581.
46. Truong T, Dang T, Phan T (2014): A study of bacterial species in cultured otter clams (*Lutraria philippinarum* Reeve, 1854) in Vietnam. Vietnam J Agric Rural Dev, **18**, 90–94.
47. Truong T, Dang T, Phan T (2015): Role of viruses (filtrated solution) in swollen-siphon disease in otter clams (*Lutraria philippinarum* Reeve, 1854) cultured in different environmental conditions. Vietnam J Agric Rural Dev, **7**, 96–101.
48. Tyson GW, Chapman J, Hugenholtz P, et al (2004): Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature, **428**, 37–43. DOI: 10.1038/nature02340.
49. Vouga M, Baud D, Greub G (2017): Simkania negevensis, an insight into the biology and clinical importance of a novel member of the Chlamydiales order. Crit Rev Microbiol, **43**, 62–80. DOI: 10.3109/1040841X.2016.1165650.
50. Zannella C, Mosca F, Mariani F, et al (2017): Microbial diseases of bivalve mollusks: infections, immunology and antimicrobial defense. Mar Drugs, **15**, 182. DOI: 10.3390/md15060182.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)