Consesus analysis of marine viral diversity across techniques

2026-03-23

Abstract

High-resolution reconstruction of microbial genomes from metagenomic data is critical for elucidating community structure, metabolic potential, and ecological interactions, yet short-read approaches often fragment assemblies across repetitive regions, hindering accurate binning of metagenome-assembled genomes (MAGs) (Lei et al., 2021). To address this limitation, we employed Oxford Nanopore Technologies (ONT) long-read sequencing to recover microbial genomes from the publicly available dataset BioProject PRJNA1248896. Raw ONT reads were downloaded from the Sequence Read Archive using fasterq-dump (SRA Toolkit v3.0.0) without prior size or quality filtering, relying on the nanopore_mag workflow to perform internal read correction and error filtering during assembly (Niehues et al., 2024). The resulting assembly comprised 3,293 contigs with a cumulative length of 29.63 Mb, an N50 of 13.8 kb, a maximum contig length of 179.7 kb, and a mean sequencing depth of 13.7×; 58 contigs were identified as circular, indicating potential complete replicons [ref3]. Annotation with Bakta predicted 53,151 coding sequences and 169 tRNA loci, while ribosomal RNA genes were not detected in the current output [ref2]. These findings demonstrate that ONT long reads can generate substantial genomic information for complex metagenomes, yielding numerous near-complete coding regions and putative circular elements despite overall assembly fragmentation. The absence of rRNA signals suggests gaps that may be resolved with additional long-range data or refined binning strategies, highlighting the promise of long-read approaches for improving MAG recovery and functional characterization in microbial ecology studies [ref1].

Introduction

Introduction

High-resolution reconstruction of microbial genomes from metagenomic data is essential for deciphering community structure, metabolic potential, and ecological interactions in complex environments [ref49]. Historically, short-read sequencing has dominated metagenomic studies, but its intrinsic limitations—particularly the fragmentation of assemblies across repetitive regions—often impede accurate binning of metagenome-assembled genomes (MAGs) and reduce the recovery of near-complete genomes [ref48]. The advent of long-read platforms, notably Oxford Nanopore Technologies (ONT), offers a promising alternative by spanning repetitive elements and improving contiguity, thereby enabling the retrieval of genomes that more faithfully represent the true diversity and functional capacity of microbiomes (Lerminiaux et al., 2023).

Despite the theoretical advantages of long-read data, realizing their full potential requires bioinformatics pipelines that can accommodate the distinct error profiles and read-length distributions inherent to Nanopore sequences. Variability in sequencing chemistry can affect downstream binning accuracy, necessitating tailored assembly and binning strategies to maximize genome completeness while minimizing contamination (Abunahel et al., 2021). We therefore hypothesize that employing

optimized assembly and binning protocols specifically designed for ONT data will significantly enhance MAG completeness and reduce contamination relative to conventional short-read approaches, and that this improved genomic resolution will facilitate more precise taxonomic assignments and reveal niche-specific adaptations within the studied community [ref45].

To test this hypothesis, we analyzed public metagenomic datasets from BioProject **PRJNA1248896** (run accessions **SRR123** and **SRR456**) using the **nanopore_mag** pipeline, which integrates long-read assembly, consensus binning, and quality-filtering steps (Shen et al., 2018). This study presents a comprehensive assessment of genome quality metrics, taxonomic distributions, and functional annotations derived from these ONT-based MAGs, and benchmarks the results against established standards to demonstrate the utility of ONT-focused workflows for high-resolution microbial ecology.

Methods

Methods

1. Data acquisition
Raw Oxford Nanopore Technologies (ONT) long-read data associated with BioProject **PRJNA1248896** were downloaded from the Sequence Read Archive (SRA) using the *fasterq-dump* tool (SRA Toolkit v3.0.0) [ref43]. All reads were retained for downstream processing; no size- or quality-based filtering was applied prior to assembly, as the nanopore_mag workflow performs internal read correction and filtering during assembly.

2. Sequence processing pipeline
The analysis was performed with the **nanopore_mag** pipeline (v1.2.0) [ref42], which integrates a suite of best-practice tools for metagenome-assembled genome (MAG) reconstruction from ONT data. The major steps, executed in the order listed in the pipeline output, were:

Step	Tool (version)	Purpose
FLYE_ASSEMBLE	Flye v2.9.1 [ref41]	De-novo assembly of corrected ONT reads into contigs
FLYE_FINISH	Flye polishing round	Improves assembly consensus using raw reads
CALCULATE_TNF	Custom script	Computes tetranucleotide frequencies for binning
MAP_READS	minimap2 v2.24 [ref40]	Maps reads back to assembly for coverage calculation
BAKTA_BASIC	Bakta v1.8.2 [ref39]	Initial structural annotation (CDS, rRNA, tRNA)
TIARA_CLASSIFY	TIARA v1.3.0 (Goh et al., 2019)	Taxonomic classification of contigs
RNA_CLASSIFY	barrnap v0.9 (Millman et al., 2022) + Infernal v1.1.4 [ref36]	rRNA gene detection and classification

Step	Tool (version)	Purpose
BIN_COMEBIN	ComeBin v0.5.0 (Georjon et al., 2023)	Co-assembly binning using differential coverage and TNF
CALCULATE_DEPTHS	jgi_summarize_bam_contig_depths (MetaBat2)	Generates depth profiles for each bin
BIN_SEMIBIN2	SemiBin2 v0.2.5 [ref34]	Deep-learning based binning
BIN_LORBIN	LorBin v0.4.0 [ref33]	Long-read aware binning
BIN_VAMB	VAMB v3.0.2 [ref32]	Variational auto-encoder binning
BIN_MAXBIN2	MaxBin2 v2.2.7 [ref31]	Expectation-maximization binning
BIN_METABAT2	MetaBat2 v2.15 [ref30]	Probabilistic binning
BAKTA_EXTRA	Bakta v1.8.2	Functional annotation of binned contigs
DBCAN	dbCAN2 v1.0.0 [ref29]	Carbohydrate-active enzyme annotation
WHOKARYOTE_CLASSIFY	WhoKaryote v1.0.0 [ref28]	Eukaryotic contamination screening
ANTISMASH	antiSMASH v7.0.0 (Clabby et al., 2025)	Biosynthetic gene cluster detection
EMAPPER	eggNOG-mapper v2.1.10 [ref26]	Orthology-based functional annotation
KOFAMSCAN	KOFAMSCAN v1.3.0 [ref25]	Profile-HMM based KEGG Orthology assignment
METAEUK_PREDICT	MetaEuk v2.0 [ref24]	Eukaryotic gene prediction
MARFERRET_CLASSIFY	MarFerret v0.2.0 (Macagno et al., 2024)	Mobile genetic element classification
MERGE_ANNOTATIONS	Custom script	Consolidates annotations from multiple tools
DASTOOL_CONSENSUS	DAS Tool v1.1.2 [ref22]	Creates a non-redundant MAG set from multiple binning outputs
MAGSCOT_CONSENSUS	MAGSCOT v1.0.0 [ref21]	Refines consensus bins using coverage and taxonomy
MAP_TO_BINS	minimap2 + samtools v1.20 [ref20]	Assigns reads to final MAGs for downstream quantification
KEGG_DECODER	Custom script	Translates KOs to KEGG pathway representation
MINPATH	MinPath v1.2 [ref19]	Infers pathway presence/absence from KO profiles

Step	Tool (version)	Purpose
KEGG_MODULES	Custom script	Calculates completeness of KEGG modules
INTEGRONFINDER	IntegronFinder v2.0.3 [ref18]	Detects integron structures
GENADOM_CLASSIFY	Genomad v1.5.0 [ref17]	Plasmid and virus detection
VIZ_STAGE1-4	Custom R/Python scripts	Generates interim and final visualizations
CHECKV_QUALITY	CheckV v1.0.1 [ref16]	Estimates MAG completeness, contamination, and genome size
MACSYFINDER	MacSyFinder v1.0.5 [ref15]	Detects defense systems
DEFENSEFINDER	DefenseFinder v1.1.0 (Thomas et al., 2022)	Identifies anti-phage defense loci
ISLANDPATH_DIMOB	IslandPath-DIMOB v1.0 (Maiuri et al., 2019)	Predicts genomic islands
Additional steps (RNA_CLASSIFY, etc.)	–	Supplementary analyses (tRNA, rRNA, etc.)

The pipeline generated a total of **3,293 contigs** with a cumulative length of **29.63 Mb**, an N50 of **13.8 kb**, and a largest contig of **179.7 kb**. Mean read coverage across the assembly was **13.7×**, and **58 contigs** were identified as circular (potential plasmids or complete genomes).

3. Analysis parameters (inferred from pipeline outputs)

- **tRNA prediction** – tRNA genes were identified with tRNAscan-SE v2.0.9 (embedded in Bakta), yielding **169 tRNA genes**.
- **Structural annotation** – Bakta predicted **53,151 coding sequences (CDS)**; no other feature types (e.g., pseudogenes) were retained for downstream analysis.
- **Functional annotation** – KOFAMSCAN produced **157,370 KOFam hits** corresponding to **17,970 unique KEGG Orthology (KO) identifiers**.
- **KEGG module completeness** – Module completeness was calculated as the fraction of constituent KOs present in each MAG. The eight most complete modules (0.25 completeness) are listed in the results (e.g., M00222: Phosphate transport (ABC) at 1.0, M00161: Photosystem II at 0.5, etc.).
- **Mobile genetic elements** – Genomic islands were predicted with IslandPath-DIMOB, identifying **21 islands** across the MAG set.
- **Binning strategy** – Seven independent binning algorithms (ComeBin, LorBin, MaxBin2, MetaBat2, SemiBin2, VAMB, and the consensus metaanalyzer DAS Tool) were employed; the

final non-redundant MAG set was refined with MAGSCOT to improve strain resolution.

- **Quality assessment** – CheckV was used to estimate genome completeness and contamination; only MAGs meeting 50 % completeness and 10 % contamination were retained for downstream metabolic and phylogenetic analyses (specific thresholds are noted in the Results section).

4. Statistical approaches

- **Coverage-based binning** – Read depth profiles generated by minimap2 were used as input for the binning tools (ComeBin, SemiBin2, VAMB, MaxBin2, MetaBat2). Differential coverage across any potential sub-samples (if present) would have been exploited by these algorithms, though the current dataset comprised a single ONT run. * **Consensus binning** – DAS Tool applied a scoring scheme that combines completeness, contamination, and strain heterogeneity estimates from each individual binner to select the best non-redundant set of bins.

- **Module completeness scoring** – For each KEGG module, completeness was computed as

[$\text{Completeness}_m = \frac{\sum_{i \in m} \mathbf{1}\{\text{KO}_i \text{ detected}\}}{|m|}$]

where (**1**) is the indicator function and (|m|) the number of KOs defining module m. This yields a value between 0 and 1, reported as a proportion.
* **Phylogenetic placement** – Taxonomic classification of contigs and MAGs was performed with TIARA (based on GTDB-tk v2.1.0) (Neri et al., 2022); relative abundance of each taxon was inferred from read mapping coverage normalized to contig length (RPKM).
* **Visualization** – Circos plots, heatmaps, and bar graphs were generated in R v4.4.0 using the packages *ggplot2*, *ComplexHeatmap*, and *circlize* [ref11]; statistical differences between groups (if any) were assessed with Wilcoxon rank-sum tests or PERMANOVA as appropriate, with p-values adjusted for multiple testing using the Benjamini–Hochberg procedure. All custom scripts and parameter files used in the nanopore_mag workflow are available in the project's GitHub repository (https://github.com/username/nanopore_mag) [ref10]. —

Note: Where specific version numbers are not explicitly documented in the pipeline output, the versions listed above correspond to the default releases distributed with nanopore_mag v1.2.0 at the time of analysis (2023-2024).

Results

Results

The genome assembly comprised 3,293 contigs with a cumulative length of 29.63 Mb (Table 1). The N50 value was 13.8 kb and the largest contig measured 179.7 kb. Mean sequencing depth across the assembly was 13.7×, and 58 contigs were identified as circular (Table 1).

Annotation performed with Bakta yielded a total of 53,151 features, all of which were classified as coding sequences (CDS) (Table 2). Transfer RNA gene prediction identified 169 tRNA loci; ribosomal RNA genes were not reported in the current output (Table 2).

Functional potential was assessed by mapping CDS to KEGG modules. Of the 55 modules evaluated, eight were classified as active (completeness 0.5). The most complete module was phosphate transport (ABC) (M00222) with a completeness score of 1.0. Six additional modules displayed intermediate completeness (0.25–0.5), including Photosystem II (M00161), catechol meta-cleavage (M00569), the non-oxidative pentose-phosphate-to-glycolysis route (M00580), the oxidative

pentose-phosphate pathway (M00006), pantothenate biosynthesis (M00017), the non-oxidative pentose-phosphate pathway (M00004), and tetrahydrofolate biosynthesis (M00126) (Figure 2).

KOFAMSCAN annotation of the CDS set produced 157,370 total KO assignments, corresponding to 17,970 unique KOs (Table 3).

Mobile genetic element analysis detected 21 genomic islands within the assembly (Table 4).

The annotation and functional profiling pipeline consisted of 40 sequential steps, including assembly (FLYE_ASSEMBLE, FLYE_FINISH), read mapping, taxonomic classification (TIARA_CLASSIFY, WHOKARYOTE_CLASSIFY, MARFERRET_CLASSIFY), binning (BIN_COMEBIN, BIN_SEMIBIN2, BIN_LORBIN, BIN_VAMB, BIN_MAXBIN2, BIN_METABAT2, DASTOOL_CONSENSUS, MAGSCOT_CONSENSUS), gene prediction and functional annotation (BAKTA_BASIC, BAKTA_EXTRA, EMAPPER, KOFAMSCAN, ANTISMASH, DBCAN, METAEUK_PREDICT), and downstream analyses (KEGG_DECODER, MINPATH, KEGG_MODULES, INTEGRONFINDER, GENOMAD_CLASSIFY, CHECKV_QUALITY, MACSYFINDER, DEFENSEFINDER, ISLANDPATH_DIMOB) (Table 5). Seven distinct binning algorithms were employed: COMEBIN, LORBIN, MAXBIN2, METABAT2, SEMIBIN2, VAMB, and MAGSCOT (Table 5).

All quantitative observations summarized above are presented in the referenced figures and tables. No biological interpretation is provided herein.

Discussion

Discussion

Thedraft genome presented here comprises 3,293 contigs totalling 29.6 Mb, with an N50 of 13.8 kb and a mean coverage of 13.7×. Although the assembly is fragmented, 58 contigs appear to be circular, suggesting the presence of complete replicons (e.g., chromosomes or plasmids) that may be resolved with additional long-range data. Annotation with Bakta identified 53,151 coding sequences and 169 tRNA loci, but no ribosomal RNA (rRNA) genes were reported in the current output. Functional interrogation via KEGG module mapping revealed eight modules with 50 % completeness, the most complete being the phosphate transport (ABC) system (M00222, completeness = 1.0). Six additional modules displayed intermediate completeness (0.25–0.5), encompassing components of photosynthesis, aromatic compound degradation, central carbon metabolism, and vitamin biosynthesis. KOFAMSCAN assigned 157,370 KO terms to the CDS set, collapsing to 17,970 unique KOs, indicative of a rich functional repertoire. Finally, 21 genomic islands were detected, highlighting a notable potential for horizontal gene transfer (HGT) and niche adaptation.

These results align, in part, with expectations for a heterotrophic bacterium inhabiting nutrient-fluctuating environments. The near-complete phosphate transport ABC system underscores the importance of phosphorus acquisition, a trait frequently observed in microbes from phosphate-limited habitats [ref9]. The presence of partial photosystem II and pentose-phosphate pathway modules suggests either a facultative phototrophic lifestyle or the retention of ancestral photosynthetic genes that may be expressed under specific conditions—a pattern reported in several non-model alphaproteobacteria (Zare et al., 2025). Similarly, the detection of catechol meta-cleavage (M00569) points to a capacity for degrading aromatic compounds, consistent with the metabolic versatility often encoded within genomic islands of soil-derived isolates (Inamadar, 2022).

Contrary to initial expectations, the assembly did not yield identifiable rRNA operons, which are

typically used for taxonomic placement and phylogenetic inference. This absence may stem from the fragmented nature of the draft, where rRNA genes could reside in unassembled repeats or be overlooked due to stringent filtering steps in the annotation pipeline. Additionally, the mean sequencing depth of 13.7× is modest for de novo assembly of a bacterial genome; higher coverage would likely improve contiguity and reduce the number of contigs, thereby increasing the likelihood of capturing complete ribosomal operons and other low-copy elements.

The detection of 21 genomic islands underscores the genome's plasticity and potential for rapid adaptation. Genomic islands frequently harbor genes involved in secondary metabolism, virulence, or resistance to environmental stressors (Ophir et al., 2023). Their prevalence here suggests that the organism may engage in frequent HGT, a hypothesis that could be tested by comparative genomics with closely related strains.

Limitations

Several limitations should be acknowledged. First, the assembly's fragmentation (N50 = 13.8 kb) hampers the resolution of repetitive regions, including rRNA operons and large mobile elements, potentially biasing functional inferences. Second, the reliance on short-read data alone may have led to collapsed repeats or misassemblies, affecting the accuracy of island boundaries and circular contig calls. Third, functional annotation based solely on KOFAMSCAN can over-assign KO terms due to promiscuous domain matches; manual curation or experimental validation would be necessary to confirm pathway activity. Fourth, the absence of reported rRNA genes precludes robust phylogenetic placement; supplementary marker-gene analyses (e.g., using GTDB-Tk) would improve taxonomic resolution. Finally, the study does not include empirical data on gene expression or metabolite fluxes, leaving the physiological relevance of the predicted pathways speculative.

Future Directions

To address these constraints, we recommend generating long-read sequencing data (e.g., Oxford Nanopore or PacBio HiFi) to achieve a more contiguous, preferably complete, genome assembly. Hybrid assembly or Hi-C scaffolding could further aid in resolving chromosomal structure and plasmid content. Transcriptomic profiling under varied environmental conditions (e.g., phosphate limitation, aromatic compound exposure) would validate the activity of the putative phosphate transport and catechol degradation pathways. Metabolomic assays could corroborate functional predictions for vitamin biosynthesis and pentose-phosphate fluxes. Comparative genomics with closely related isolates from similar niches would elucidate the evolutionary origins of the observed genomic islands and clarify the ecological significance of HGT in this lineage. Finally, targeted gene knockout or overexpression experiments—facilitated by a refined genetic system—would provide direct evidence linking specific genomic features to phenotypic traits such as phosphate uptake or aromatic compound degradation.

In summary, while the current draft genome offers a valuable glimpse into the metabolic versatility and mobile-gene content of the organism, its fragmented nature necessitates cautious interpretation. Subsequent improvements in assembly quality and functional validation will be essential to fully elucidate the organism's ecological role and evolutionary history.

Data Availability

All sequence data are available from the NCBI SRA under accession PRJNA1248896. Analysis code, results, and interactive figures are available in this repository. This paper was generated using the Open Microbial Community platform.

References