CompleteBin: A transformer-based framework unlocks microbial dark matter through improved short contig binning

Bohao Zou[1†], Zhenmiao Zhang[1,2†], Xiaohan Wang[3,4†],
Rong Tao[1], Nianzhen Gu[1], Karsten Kristiansen[4],
Mo Han[3,5,6*], and Lu Zhang[1,7*]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China
[2]Department of Computer Science & Engineering, UC San Diego, CA 92122, USA
[3]BGI Research, Sanya 572025, China
[4]Laboratory of Integrative Biomedicine, Department of Biology, University of Copenhagen, 2100 Copenhagen, Denmark
[5]State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen 518083, China
[6]Shenzhen Key Laboratory of Environmental Microbial Genomics and Application, BGI Research, Shenzhen 518083, China
[7]Institute of Systems Medicine and Health Sciences, Hong Kong Baptist University, Hong Kong
[†]These authors contributed equally to this work
[*]To whom correspondence should be addressed: hanmo@genomics.cn, ericluzhang@hkbu.edu.hk

Abstract

Metagenomic binning is crucial for reconstructing microbial genomes from metagenomic sequencing samples. However, existing tools struggle in complex communities where short, low-abundance contigs predominate, thereby limiting the recovery of complete metagenome-assembled genomes (MAGs) and the identification of novel functions. Here, we introduce CompleteBin, a Transformer-based framework that integrates contig sequence context, pre-trained taxonomic embeddings from a genome language model, and dynamic contrastive learning to bin short contigs robustly. Across CAMI II datasets, CompleteBin increased near-complete MAG recovery by 38.5% over leading methods like COMEBin. Across diverse real-world datasets (marine, freshwater, plant-associated, cold seep sediment, and human gut), it achieved a 57.4% improvement on average. Applying CompleteBin to six cold seep sediment samples uncovered 129 strain-level genome bins across 30 phyla, including 13 phyla undetected by other tools, and taxonomically assigned 90,405 genes (32.1% of total), revealing previously unknown species in nitrogen and sulfur cycling. CompleteBin unlocks microbial dark matter in diverse environments, advancing our understanding of microbial ecology and biogeochemical processes.

Introduction

Metagenomic sequencing has fundamentally transformed our understanding of microbial diversity by enabling culture-independent exploration of environmental microbiomes, revealing vast numbers of previously unknown taxa and their ecological functions [1, 2, 3, 4]. The reconstruction of metagenome-assembled genomes (MAGs) through metagenomic binning, a computational process that clusters assembled contigs derived from the same microbial genome, represents a cornerstone methodology for linking phylogenetic identity to metabolic potential [5]. The quality and comprehensiveness of binning results critically influence all downstream analyses, from taxonomic profiling to functional annotation, making the development of robust and efficient binning methods a paramount research priority.

A fundamental and largely unresolved computational challenge in metagenomic binning lies in the effective processing of short contigs ($\leq$2,500 bp). Despite significant algorithmic advances in recent years [6], these short contigs remain systematically difficult to classify accurately. Yet, they represent critically important sources of genetic information for high-quality MAG reconstruction, particularly in highly diverse microbial communities [7]. Our analysis of diverse metagenomes reveals that short contigs contribute disproportionately more nucleotide bases as community diversity increases (**Figure 1a**; linear regression, $p = 1.54 \times 10^{-4}$) and harbor significantly more complete protein-coding genes (**Figure 1b**; linear regression, $p = 3.52 \times 10^{-4}$). Conversely, the relative information content from long contigs (>2,500 bp) shows a negative correlation with community diversity (**Figure 1a-b**;

Supplementary Note 1), indicating a systematic shift in metagenome assembly characteristics as microbial complexity increases. Additionally, we observed that short contigs are predominantly derived from low-abundance microbial species that face assembly challenges due to insufficient sequencing coverage ($p < 2.2 \times 10^{-16}$, **Figure 1c**; **Supplementary Note 2**), yet may play disproportionately important roles in ecosystem functioning [8].

Current metagenomic binning tools address short contigs through two strategies, each with significant methodological limitations. The first approach, implemented by MetaBAT2[9], MetaDecoder[10], and VAMB[11], completely excludes short contigs from MAG reconstruction. While this conservative strategy simplifies the computational problem and reduces potential contamination, it systematically discards potentially complete protein-coding genes and low-abundance species, making it inherently suboptimal for comprehensive microbial genome recovery and ecological understanding. The second strategy, adopted by MaxBin2[12], CONCOCT[13], SemiBin2[14], and COMEBin[6], attempts to incorporate short contigs (typically 1,000-2,500 bp) into the binning process but relies heavily on sequencing coverage and tetranucleotide frequency (TNF) signatures. This approach faces fundamental technical limitations that become increasingly ineffective as contigs get shorter. Our analysis reveals that TNF specificity between two species is lost for short contigs (**Figure 1d, e**; **Supplementary Note 3**). Additionally, contig sequencing coverage is well known to be less effective for single-sample binning [15]. These fundamental limitations create a critical gap in current methodologies, highlighting the urgent need for approaches that can accurately handle short contigs to maximize microbial genome recovery from complex metagenomic samples.

To address these fundamental challenges, we developed CompleteBin, a novel deep learning framework specifically engineered to overcome the computational bottlenecks associated with short contig binning through three synergistic methodological innovations. 1. Context-Based TNF Tokenization: unlike previous methods that rely solely on global TNF vectors for contig representation, CompleteBin employs a context-aware tokenization strategy that concatenates global and local TNF vectors into sequential representations. This approach captures both compositional signatures and local sequence context, preserving crucial positional information that becomes particularly important for maintaining specificity in short contigs (**Figure 1h-i**). 2. Pre-trained deep language model: CompleteBin incorporates external taxonomic knowledge through a transformer model pre-trained on 15,780 publicly available microbial reference genomes and their corresponding species-level taxonomic lineages. This pre-training approach creates taxonomic embeddings that implicitly encode vast phylogenetic relationships without requiring computationally expensive alignment procedures during binning (**Figure 1i**). The pre-trained model links genomic sequences to taxonomic classifications, effectively compensating for TNF instability in short contigs by leveraging evolutionary relationships learned from

comprehensive genomic databases. 3. Dynamic Multi-View Contrastive Learning: to enable unified representation of contigs with dramatically different lengths, CompleteBin implements a sophisticated contrastive learning framework that dynamically generates multiple subsequence views for each contig during training. This approach minimizes distances between views derived from the same contig while maximizing separation between views from different organisms (**Figure 1j**; **Methods**). The dynamic generation strategy provides diverse training examples that significantly improve embedding quality for short contigs, while self-dropout masking promotes consistency under varying model conditions. CompleteBin employs a two-stage clustering process that combines Leiden [16] algorithm-based initial clustering with FLSpp [17] refinement to minimize contamination while maximizing genome recovery (**Figure 1k**; **Methods**).

We evaluated CompleteBin's performance in metagenomic binning by integrating short contigs across datasets with varying complexity (**Figure 2**). Our results demonstrate that CompleteBin's performance advantage increases proportionally with microbial community complexity on both CAMI II [18] (**Figure 2a**) and real-world datasets (**Figure 2d**). We further showed that short contigs contribute essential microbial marker genes that could improve MAG completeness and quality (**Figure 2g** and **2h**). Comparative analysis with the current state-of-the-art method COMEBin [19] reveals that CompleteBin tends to recover more near-complete MAGs (NCMAGs; completeness $\geq$ 90% and contamination $\leq$5%) and medium-quality MAGs (MQMAGs, completeness $\geq$ 50% and contamination $\leq$ 10%, **Methods**) of low-abundance genomes on CAMI II marine dataset (**Supplementary Figure 2**). Next, we comprehensively benchmarked CompleteBin against the existing metagenomic binning tools, including CONCOCT, MetaBAT2, SemiBin2, VAMB, and COMEBin, on both CAMI II (human microbiome, marine, and plant) and diverse real-world datasets (freshwater, marine, cold seep sediment, plant-associated, and human gut microbiomes). CompleteBin consistently outperforms the other methods in recovering NCMAGs and MQMAGs. Specifically, CompleteBin increases the number of NCMAGs by 38.5% across CAMI II datasets and by 57.4% across real-world samples compared to the second-best method (**Figure 3-5**). We further assessed the practical application of CompleteBin by analyzing its performance on six metagenomic samples of cold seep sediments (**Methods**). CompleteBin identifies 129 near-complete strain-level genome bins (StGBs, **Methods**) spanning 30 phyla that were completely unrecovered by the other binning tools, where 13 phyla were detected exclusively by CompleteBin, particularly the small-genome archaeal phylum Nanoarchaeota. CompleteBin exclusively assigned 90,405 novel genes (32.1% of total identified genes) with diverse metabolic functions to taxonomically resolved StGBs, substantially expanding our understanding of microbial diversity and its biogeochemical processes in these extreme environments (**Figure 6**).

Results

Architecture of CompleteBin

CompleteBin is a Transformer-based deep learning framework for metagenomic binning that can effectively process both long and short contigs (≥768 bp; **Methods**). The architecture integrates three fundamental information sources for comprehensive contig representation: context-based TNF tokenization, sequencing coverage embeddings, and taxonomic embeddings derived from pre-trained genome language models (**Figure 1i**). The context-based TNF tokenization system divides each contig into a fixed number of equal-length fragments (16 fragments by default), computing TNF for both the entire contig (global TNF) and individual fragments (local TNFs). These signatures are concatenated sequentially to capture both compositional patterns and local sequence context, along with their positional relationships in the contig (**Figure 1h**). This approach provides stable sequence context information that remains informative even for short contigs, where traditional global TNF features lose specificity. Sequencing coverage embeddings are generated through multi-layer perceptrons (MLPs) that process the mean and standard deviation of contig coverage, providing abundance information that complements compositional features. Taxonomic embeddings are derived from a pre-trained multi-modal Transformer model, trained on 15,780 publicly available microbial reference genomes and species-level taxonomic lineages (**Methods**). This pre-training procedure links genomic sequences to taxonomic lineages, mitigating TNF instability for short contigs.

CompleteBin employs dynamic multi-view contrastive learning and self-dropout masking to produce robust contig embeddings (**Figure 1j**; **Methods**). The dynamic multi-view approach generates multiple subsequence views for each contig during training, with the optimization objective encouraging similarity between views from the same contig while maximizing dissimilarity between views from different organisms. This strategy provides more diverse subsequences for model training, which could significantly enhance model robustness and substantially reduce the risk of overfitting. Self-dropout masking promotes consistency by encouraging similar embeddings for identical contigs under varying dropout conditions, improving model robustness and generalization. The final clustering process implements a two-stage approach (**Figure 1k**; **Methods**). Initial clustering employs the Leiden algorithm with diverse hyperparameter configurations, considering both contig embeddings and single-copy gene constraints. High-contamination clusters undergo refinement using the FLSpp algorithm, with final ensemble results generated through Galah to optimize both purity and completeness across multiple clustering solutions.

Performance scaling with microbial community complexity

High microbial diversity substantially complicated genome assembly processes, resulting in increased production of short contigs. To systematically assess performance scaling of CompleteBin, we quantified microbial diversity across CAMI II datasets (human microbiome, marine, and plant environments; total 80 samples) and real-world samples (freshwater, marine, plant-associated, and human fecal microbiomes; total 10 samples) using established diversity metrics (**Methods**). Relative increase of NCMAG between CompleteBin and COMEBin (current best binning method [19]) revealed strong positive correlations with diversity levels in both CAMI II (linear regression $R^2 = 0.42$, $p = 5.01 \times 10^{-5}$; **Figure 2a**) and real-world datasets ($R^2 = 0.38$, $p = 7.54 \times 10^{-3}$; **Figure 2b**). These results demonstrate that CompleteBin achieves maximum benefit precisely in the complex communities where comprehensive microbial characterization is most challenging yet scientifically valuable.

Integration of short contigs reveals low-abundance microbial genomes

Low-abundance microorganisms typically suffered from fragmented assemblies due to insufficient sequencing coverage [7], resulting in numerous short contigs ($p < 2.2 \times 10^{-16}$, **Supplementary Note 4**). To quantify the specific contributions of short contigs to genome recovery and MAG quality enhancement, we systematically compared CompleteBin performance using long contigs only (>2,500 bp, Bin_long) versus combined short and long contigs (≥900 bp, Bin_short) on the CAMI II marine dataset. For the total 977 reference genomes in this dataset, Bin_short uniquely recovered 185 genomes compared to only 8 unique genomes identified exclusively by Bin_long. The genomes uniquely recovered by Bin_short exhibited significantly lower abundance than the genomes detected by both settings (Wilcoxon rank-sum test, $p < 2.2 \times 10^{-16}$, **Figure 2e**). The MAGs from these unique genomes demonstrated significantly reduced N50 values compared to the MAGs from genomes intersecting between both settings (Wilcoxon rank-sum test, $p < 2.2 \times 10^{-16}$ for both, **Figure 2f**), indicating that short contig integration specifically enables recovery of fragmented genomes from non-abundant taxa that would otherwise remain undetected.

Beyond discovering novel genomes, short contig integration substantially improved the quality of shared MAGs. Our analysis revealed that Bin_short enhanced the quality of 147 shared MAGs, upgrading low-quality MAGs to MQMAGs or MQMAGs to NCMAGs. The reference genomes corresponding to these 147 upgraded MAGs showed significantly lower abundance compared to other MQMAGs and NCMAGs from Bin_long (Wilcoxon rank-sum test, $p < 2.2 \times 10^{-16}$; **Figure 2c**). Assembly contiguity analysis demonstrated that upgraded MAGs in Bin_short exhibited significantly lower N50 values than other MQMAGs and NCMAGs in Bin_long (Wilcoxon rank-sum test, $p < 2.2 \times 10^{-16}$; **Figure 2d**), suggesting that quality improvements resulted from incorporating short contigs to fill assembly

gaps. We provided two illustrative examples demonstrating how short contig recruitment transforms low-quality MAGs to MQMAGs and MQMAGs to NCMAGs (**Figure 2g-h**). The functional impact of short contig integration proved substantial. Upgraded MAGs in Bin_short contained 282,365 genes compared to 194,915 genes in their corresponding Bin_long counterparts, representing a 44.9% increase in gene recovery. This dramatic improvement in complete gene identification highlights the critical importance of short contigs for comprehensive functional annotation, taxonomic annotation, and other downstream genomic analyses. We benchmarked CompeteBin against COMEBin using contigs with a minimum length of 1,000 bp. This analysis further revealed that CompeteBin outperformed COMEBin in both low-abundance genome recovery and overall MAG quality by more effectively handling short contigs (**Supplementary Note 5**).

Benchmarking demonstrates CompleteBin superiority across diverse metagenomes with single-sample binning mode

We conducted a systematic evaluation of CompleteBin against existing methods (CONCOCT, MetaBAT2, SemiBin2, COMEBin) across 80 metagenomes from three CAMI II datasets (human microbiome, marine, and plant) and 16 metagenomes from five real-world environments. CompleteBin consistently outperformed all competing methods across the three CAMI II datasets. In the human microbiome dataset, which covers Gastrointestinal, Urogenital, Airways, Skin, and Oral samples (total 49 samples), CompleteBin recovered substantially more NCMAGs than any other approach, with increases of 271 (76.1%), 212 (51.1%), 185 (41.8%), and 96 (18.1%) NCMAGs relative to CONCOCT, MetaBAT2, SemiBin2, and COMEBin, respectively (**Figure 3a**). This pattern was also observed in the marine (10 samples) and plant (21 samples) datasets, with CompleteBin achieving aggregate improvements of 38.5% over the best alternative method (**Figure 3b-c**). Additionally, CompleteBin consistently generated the highest number of high-purity MAGs with contamination levels below 10% for all three CAMI II datasets (**Figure 3d-f**). F1 score (bp) analysis confirmed the robustness of CompleteBin, showing that it consistently produced the highest number of MAGs exceeding all evaluated thresholds from 0.5 to 0.9 (**Methods**; **Supplementary Figure 3, 4, and 5**).

Real-world metagenomes revealed even greater performance gaps between CompleteBin and other tools. Across 16 diverse samples, comprising marine (3 samples) [20], freshwater (3 samples) [21, 22, 23], plant-associated (1 sample) [24], cold seep sediment (6 samples) [25], and human gut microbiomes (3 samples) [26, 27], CompleteBin generated average improvements of 57.4% more NCMAGs than the second-best method (**Figure 4**). Across four distinct microbiome categories (marine, freshwater, plant-associated, and human gut), CompleteBin recovered a substantially greater number of NCMAGs compared to existing binning methods. On average, CompleteBin outperformed CONCOCT, MetaBAT2,

SemiBin2, and COMEBin by 449 (1,360.6%), 326 (208.9%), 269 (126.3%), and 165 (52.1%) NCMAGs, respectively (**Figure 4g-p**). The performance advantage of CompleteBin was most pronounced in high-complexity samples in cold seep sediment. Specifically, CompleteBin recovered 116 NCMAGs compared to 63 NCMAGs obtained by COMEBin, representing an 84.1% improvement over the best alternative method across all cold seep sediment samples (**Figure 4a-f**). Furthermore, CompleteBin successfully produced a total of 1,415 MAGs above medium-quality criteria, which far exceeded the 134, 430, 465, and 414 MAGs above medium-quality MAG criteria generated by CONCOCT, MetaBAT2, SemiBin2, and COMEBin, respectively (**Supplementary Figure 6**). We further evaluated the mapping ratio (proportion of total reads that are mapped to MAGs, **Methods**) for the NCMAGs and MQMAGs generated by each binning tool. Across 16 diverse samples, CompleteBin achieved an average improvement of 50.12% higher mapping ratio than the second-best method (**Figure 4**). These findings suggest that CompleteBin captured more of the organism's genome sequences and identified additional genomes from the microbiomes that other binning tools failed to detect.

While CompleteBin's performance advantages positively correlated with microbial diversity, the relationship was not strictly linear, as evidenced by the CAMI II marine and plant datasets. Compared to the second-best performing method, CompleteBin achieved a 29.7% improvement for the marine dataset (average diversity: 20.6) and a 47.2% enhancement for the plant dataset (average diversity: 19.2). The higher average diversity in the marine dataset corresponded to a lower percentage improvement compared to the plant dataset. This apparent deviation may be attributed to differences in assembly quality between the datasets. The marine dataset exhibited substantially shorter contigs (average N50: 1,150 bp) compared to the plant dataset (average N50: 2,112 bp), suggesting that improved assembly contiguity may enhance CompleteBin's capacity for NCMAG recovery, potentially compensating for or amplifying the effects of community diversity.

Multi-sample binning mode amplifies performance advantages

To evaluate the multi-sample binning (**Methods**) capabilities of different binning tools, we assessed six methods (CONCOCT, MetaBAT2, VAMB, SemiBin2, COMEBin, and CompleteBin) on three comprehensive datasets: freshwater samples (3 samples), cold seep sediments (6 samples), and CAMI II marine samples (10 samples). CompleteBin demonstrated exceptional multi-sample performance across all datasets. On the CAMI II marine dataset, it recovered 81 (49.3%), 125 (104.2%), 57 (30.3%), 93 (61.1%), and 58 (31.0%) more NCMAGs than VAMB, CONCOCT, MetaBAT2, SemiBin2, and COMEBin, respectively (**Supplementary Figure 8**). F1 score (bp) analysis confirmed superior performance of CompleteBin across quality thresholds between 0.5 and 0.8 (**Supplementary Figure 10**).

Freshwater dataset analysis revealed remarkable improvements, with CompleteBin recovering 141 (2,350%), 121 (605%), 99 (206%), 91 (163%), and 48 (49%) additional NCMAGs compared to VAMB, CONCOCT, MetaBAT2, SemiBin2, and COMEBin, respectively (**Figure 5a-c**). Cold seep sediment analysis showed similarly impressive gains: 200 (952%), 192 (662%), 169 (325%), 173 (360%), and 88 (66%) more NCMAGs than the VAMB, CONCOCT, MetaBAT2, SemiBin2, and COMEBin (**Figure 5d-i**). The recovery of MQMAGs demonstrated consistent superiority across all multi-sample datasets (**Supplementary Figures 7, 11**). These results establish CompleteBin's multi-sample binning mode as significantly superior to current approaches, with particularly dramatic improvements observed in complex environmental samples.

The efficacy of multi-sample binning relative to single-sample binning varied substantially between freshwater and cold seep sediment datasets across all binning tools. For the freshwater samples, CompleteBin demonstrated only marginal improvement in multi-sample mode, yielding a modest 3.5% increase in the number of NCMAGs compared to single-sample binning. This limited improvement was consistent with other binning tools, where SemiBin2 showed a 5.1% decrease and COMEBin exhibited a 6.4% increase. In contrast, the cold seep sediment samples revealed dramatically enhanced performance with multi-sample binning, where CompleteBin generated 90.5% more NCMAGs compared to single-sample mode. Similar improvements were observed for SemiBin2 (45.4% increase) and COMEBin (111.1% increase). Two factors likely contributed to this differential performance between datasets. First, the freshwater dataset comprised only three samples for multi-sample binning compared to six samples in the cold seep sediment dataset, suggesting that larger sample sizes may enhance multi-sample binning efficacy. Second, substantial differences in contig coverage patterns were observed between the two datasets. The freshwater samples exhibited poor cross-sample coverage, with an average of 48.7% of contigs lacking coverage values from at least one sample. Conversely, the cold seep sediment samples demonstrated superior consistency in cross-sample coverage, with an average of only 15.7% of contigs missing coverage values from other samples. These findings indicate that the selection of metagenomic samples with compatible coverage profiles is critical for optimizing multi-sample binning performance and maximizing NCMAG recovery.

Deciphering cold seep sediment microbiome and their metabolic networks using CompleteBin

CompleteBin identified 129 near-complete StGBs spanning 30 phyla that remained completely unrecovered by conventional binning tools. These discoveries represent a substantial expansion of known microbial diversity in cold seep ecosystems, with taxonomic assignments including Bacteroidota (25 genomes) and Desulfobacterota (41 genomes), which play essential roles in nitrogen and sulfur biogeo-

chemical cycling. Critically, 13 phyla were exclusively detected by CompleteBin, including Nanoarchaeota, an archaeal lineage characterized by exceptionally small genome sizes that pose particular challenges for conventional assembling and binning approaches. Among the 23 identified Patescibacteria StGBs, representing candidate phyla radiation bacteria also with reduced metabolic capabilities, 14 were exclusively recovered by CompleteBin. Additionally, four StGBs were not taxonomically classifiable at the phylum level, suggesting the presence of potentially novel high-level taxonomic groups (**Figure 6a**; **Supplementary Table 1**).

Comprehensive functional analyses were conducted for all 281,392 non-redundant complete genes recovered by CompleteBin, of which 90,405 (32.1%) could not be taxonomically traced by other binning tools. These novel genes demonstrate diverse functional capabilities, with 69,130 receiving annotations (**Methods**). Functional categorization identified 20,950 genes related to metabolism, 12,987 involved in cellular processes and signaling, and 11,598 associated with information storage and processing, with only 5,480 genes remaining functionally uncharacterized. Enzyme classification analysis revealed that most proteins encoded by these genes represent novel catalytic functions not included in existing databases. Among the 26.3% of genes encoding proteins with Enzyme Commission number assignments, the distribution included 6,608 transferases, 3,720 oxidoreductases, 3,714 hydrolases, 1,509 lyases, 1,188 isomerases, and 1,430 ligases (**Figure 6b**; **Supplementary Table 2**), indicating substantial expansion of known enzymatic diversity in extreme environments.

Community-scale metabolic network modeling provided deeper insights into the functional significance of CompleteBin's unique discoveries. While the total number of metabolic substrates did not show dramatic increases, we identified substantially more enzymes, transporters, and specialized reactions, indicating that previously undetected organisms contribute significantly to microbial metabolism and inter-species interactions in cold seep sediments (**Figure 6c**). Detailed analysis of nitrogen and sulfur metabolism revealed unprecedented metabolic versatility among CompleteBin unique genomes. We identified 22 StGBs with complete nitrate reduction pathways (NO_3^- to NH_3, **Figure 6d**) and 34 StGBs capable of complete sulfate reduction (SO_4^{2-} to H_2S, **Figure 6e**). Remarkably, 20 StGBs demonstrated functionality in both pathways, spanning five major phyla, including Desulfobacterota, Bacteroidota, Pseudomonadota, Acidobacteriota, and Chloroflexota. These dual-functional organisms highlight previously unrecognized connections between nitrate and sulfur cycling in cold seep ecosystems. Their metabolic flexibility may resolve longstanding questions about energy limitations in sulfate-dependent anaerobic oxidation of methane (AOM) while providing alternative electron acceptor utilization under fluctuating geochemical conditions. The ability to switch between nitrate and sulfate reduction represents a critical adaptational advantage for maintaining metabolic resilience under fluctuating geochemical conditions in cold seep sediments.

Discussion

The systematic exclusion or misassignment of short contigs represents a fundamental limitation that impedes comprehensive microbial genome recovery. Our analysis reveals that short contigs contain substantial biological information, including complete protein-coding genes related to essential metabolic functions, cellular processes, and signaling pathways. Critically, these short contigs often originate from low-abundance species that may exert disproportionately large ecological influence. Existing binning methodologies either discard short contigs entirely or assign them incorrectly due to unreliable coverage profiles and TNF features, thereby creating substantial gaps in genome reconstruction for low-abundance taxa. This limitation becomes particularly pronounced in complex environmental samples, where assembly processes generate numerous short contigs and may lead to an even more problematic situation in that some of the essential species may be completely lost. To address these challenges, CompleteBin incorporates several methodological advances, including context-based TNF tokenization, advanced model architecture with pre-trained deep language models, and dynamic contrastive learning. Evaluation across a total of 96 CAMI II and real-world metagenomic samples demonstrates that CompleteBin consistently achieves superior binning accuracy and completeness compared to existing methods, particularly for recovering genomes from low-abundance microbial communities that are often overlooked by conventional approaches.

Short contigs are particularly prevalent in deciphering complex environmental metagenomes using genome assembly. Recent analysis [7] demonstrates that soil samples require 0.9-4.6 terabases of clean reads per sample to achieve near-complete metagenomic coverage, over 1,500-fold more than the sequencing depth required for human gut samples. Such sequencing depths remain prohibitively expensive and time-consuming to achieve with current technologies. Critically, insufficient sequencing coverage directly compromises assembly quality, as evidenced by strong positive correlations between metagenomic depth and both assembly N50 values and read recruitment rates. Even at ultra-deep metagenomic sequencing exceeding 500 Gb, soil metagenome assemblies recruit only 27-57% of reads to contigs longer than 2,000 bp, inevitably yielding substantial proportions of fragmented assemblies dominated by short contigs. In highly diverse environmental microbiomes, assembly fragmentation remains largely unavoidable given practical sequencing and computational constraints. CompleteBin addresses this limitation through context-based TNF tokenization, which preserves compositional information in short sequences, and a pre-trained language model that compensates for the coverage limitations inherent to fragmented assemblies. The substantial performance improvements observed in complex real-world samples, including an 84.1% improvement in the number of NCMAGs over the second-best binning method in cold seep sediments. This suggests that advancing binning algorithms to effectively process short contigs may represent a more tractable solution than pursuing the prohibitive sequencing

depths required for high-quality de novo assembly in these challenging environments.

While SemiBin2 and COMEBin use contrastive learning for binning, CompleteBin addresses several key limitations. First, both methods construct augmented data before training, limiting data diversity, whereas CompleteBin uses dynamic augmentation during training to learn diverse subsequence patterns and reduce overfitting. Second, their reliance on TNF features alone results in information loss for short contigs, whereas CompleteBin's context-based TNF tokenization preserves both global and local patterns across variable contig lengths. Third, CompleteBin's Transformer architecture better captures sequential dependencies than their MLP-based models. Finally, CompleteBin integrates taxonomic information via pre-trained embeddings without requiring the time-consuming alignments needed by SemiBin1, unlike SemiBin2 and COMEBin, which ignore taxonomy entirely.

Long-read sequencing presents a promising alternative to short-read technologies for generating high-quality microbial genomes. However, its utility in complex environmental samples, such as soil microbiomes, remains limited by significant technical challenges [28]. Key obstacles include the difficulty of obtaining high-quality, high-molecular-weight DNA, the vulnerability of this DNA to damage and contamination [29], and the high cost associated with sequencing low-abundance taxa to adequate coverage in highly diverse communities. Consequently, short-read sequencing remains the predominant approach for characterizing complex environmental microbiomes. CompleteBin was developed specifically to maximize binning accuracy and completeness with the short-read data. Our results demonstrate that CompleteBin's performance advantage over existing methods becomes increasingly pronounced as microbiome diversity increases. This scaling behavior makes CompleteBin particularly advantageous for analyzing complex environmental samples, where short-read sequencing technologies continue to represent the most practical experimental approach.

A limitation of CompleteBin is its performance on low-complexity datasets with fewer than 8,162 contigs, which provide too few training steps per epoch for stable learning (**Supplementary Note 6**). We note that such datasets, often resulting from high-quality assemblies, are generally easy to recover microbial genomes. This limitation could also be addressed within CompleteBin by implementing a repeat sampling strategy in training.

Methods

Overview of CompleteBin

CompleteBin is a Transformer-based deep learning framework for metagenomic binning of contigs with lengths $\geq$768 bp. The framework operates through four stages (**Figure 1**). In the first stage, contigs and their augmented views undergo context-based TNF tokenization. The tokenized sequences

are processed through two parallel pathways: a Transformer model for generating dynamic contig representations and a pre-trained Transformer model for extracting taxonomic embeddings. Contig coverage features (mean and standard deviation) are processed through an MLP. In the second stage, a fusion MLP integrates the taxonomic embeddings, dynamic contig representations, and abundance features to produce unified contig embeddings. We employ dynamic multi-view contrastive learning coupled with self-dropout masking strategies for training robust contig embeddings. In the third stage, initial clustering is performed using the Leiden community detection algorithm with the contig embeddings. Subsequently, high-contamination clusters are refined using the FLSpp algorithm. In the final stage, we apply Galah to consolidate clustering results obtained across multiple hyperparameter configurations, thereby optimizing the trade-off between MAG completeness and contamination levels.

Context-based TNF tokenization and abundance of contigs

The TNF vector was generated using a sliding window of size k with a stride of one to traverse the sequence, producing multiple k-mer fragments. The frequency distribution of these k-mer fragments was calculated, resulting in a vector with T dimensions. We set k to 4 and treated reverse complement sequences as equivalent. Consequently, the TNF vector had 136 dimensions ($T = 136$). The exact formulas used in the calculation are provided below.

$$F' = (f_1, ..., f_j, ..., f_{136}) \tag{1}$$

where f_j represents the frequency of the j-th k-mer fragment. We further normalized the frequency distribution F' with

$$q_j = \frac{f_j}{\sum_{i=1}^{136} f_i} \tag{2}$$

Finally, the TNF vector could be calculated as

$$F = (q_1, ..., q_j, ...q_{136}) \tag{3}$$

To calculate the local TNF vector, the contig was segmented into a fixed number of patches with equal sequence length. The local TNF vector was calculated separately for each patch to capture local compositional patterns. These local TNF vectors were concatenated in their original sequence order, preserving both local variations and the broader structural context of the sequence. Furthermore, a global TNF vector was added at the beginning of the ordered local TNF vectors. The context-based TNF tokenization for a contig could be calculated as follows:

$$S = [F_{global}, F_{p_1}, ..., F_{p_j}, ..., F_{p_l}] \tag{4}$$

where F_{global} and F_{p_j} represent the global TNF vector of the whole contig and the local TNF vector of the j-th patch, respectively. l is the number of patches. We set l as 16 in our study.

Coverage features for contigs were derived following the methodology established in COMEBin. For each contig, the mean and standard deviation of read coverage depth were calculated across all nucleotide positions. For multi-sample binning, coverage features from individual sequencing samples were concatenated to generate composite feature vectors for each contig. To normalize for variation in sequencing depth across samples, we applied the normalization procedure described in COMEBin, thereby correcting for potential biases arising from differential read abundances between samples.

Architecture of the Transformer and MLPs in CompleteBin

The Transformer architecture consisted of three Transformer encoder layers [30] with each layer containing 512 hidden dimensions. The feedforward neural networks within each Transformer encoder layer employed sigmoid linear units (SiLU) [31] as the activation function. Root mean square layer normalization [32] was applied throughout the architecture, and dropout regularization (p = 0.15) [33] was implemented in both the multi-head attention mechanisms and feedforward neural networks of each Transformer encoder layer.

The abundance MLP module consisted of three fully connected layers with progressive dimensionality reduction. The first, second, and third layers contained 136, 2,048, and 1,024 hidden dimensions, respectively. Rectified linear units (ReLU) [34] served as the activation function throughout this module. Batch normalization [35] and dropout regularization (p = 0.15) were applied to prevent overfitting and improve training stability. The fusion MLP module employed a six-layer fully connected architecture with systematic dimensionality reduction across layers. Hidden dimensions were configured as 2,560, 4,096, 3,072, 2,048, 1,024, and 100 for the first through sixth layers, respectively. Similar to the abundance MLP module, ReLU activation functions were utilized, with batch normalization and dropout regularization (p = 0.15) applied consistently throughout the network.

All model components were implemented using the PyTorch framework (v2.1.0). GPU acceleration was enabled through the CUDA toolkit (v11.4).

Pre-train the Transformer model with reference genomes and taxonomic lineages

Reference genomes of microbiomes were obtained from the proGenomes v3.0 database [36], with corresponding complete taxonomic lineages retrieved from the National Center for Biotechnology Information (NCBI). Quality control was implemented through a multi-step filtering process. Initially, the number of species within each phylum was quantified, and phyla containing fewer than 10 species were excluded from subsequent analyses. Redundant entries with identical taxonomy IDs were removed, with preference given to the genome of maximum length. Additionally, we only retained the largest genome per species when multiple strains were present. This systematic filtering procedure yielded a curated dataset comprising 15,780 distinct representative genomes, each corresponding to a unique microbial species.

A Transformer model with three Transformer encoder layers was employed for pre-training on the curated genomic dataset. The pre-training methodology utilized stochastic sampling of sequences from the reference genomes, with sequence lengths uniformly distributed across a range of 768 to 60,000 base pairs. Input sequences were tokenized using context-based TNF tokenization. The training objective was formulated as a multi-class classification task, wherein each reference genome's species-level taxonomic annotation constituted a discrete class label, resulting in 15,780 distinct classification categories. Model optimization was conducted over 128 training epochs with a batch size of 256. The training procedure employed the AdamW optimizer with an initial learning rate of 1×10^{-5}, subject to cosine annealing scheduling to facilitate convergence.

Dynamic contrastive learning framework

Dynamic data augmentation

CompleteBin employed the data augmentation strategy during the training process rather than pre-generating augmented data (SemiBin2, COMEBin). At each training iteration, a batch of contigs (default batch size: 1,024) was collected from the dataset, followed by real-time application of data augmentation to generate contrastive pairs. The model was subsequently trained on both the original contigs and their corresponding augmented pairs within the same training step. This procedure was iteratively executed throughout each training epoch. The stochasticity of the data augmentation process would result in identical contigs generating distinct augmented counterparts across different epochs.

Dynamic multi-view contrastive learning

CompleteBin dynamically generated multi-views (subsequences) for each contig in a batch during the whole training stage. This optimization goal was designed to pull the representations of a contig and its generated subsequences closer while maintaining the distance between the representations of other contigs and their subsequences within the same batch. The corresponding loss function for two subsequences (L_{ts}) could be expressed as follows:

$$\ell(\theta_{i,s_1}, \theta_{i,s_2}) = \frac{exp(cos(\theta_{i,s_1}, \theta_{i,s_2})/\tau)}{\sum_{b=1}^{B} 1_{b\neq i}[\sum_{j=1}^{2} exp(cos(\theta_{i,s_1}, \theta_{b,s_j})/\tau)] + exp(cos(\theta_{i,s_1}, \theta_{i,s_2})/\tau)} \tag{5}$$

$$L_{ts} = -log(\ell(\theta_{i,s_1}, \theta_{i,s_2})) \tag{6}$$

where θ_{i,s_1} and θ_{i,s_2} are the representations of the two subsequences of the i-th contig. B represents the batch size. $cos(\cdot,\cdot)$ is the cosine similarity between two vectors, and τ is the temperature.

We expanded the formula from two subsequences to multiple subsequences. The final dynamic multi-view contrastive loss (L_{dmv}) could be calculated as follows:

$$L_{dmv} = \frac{1}{BM(M-1)} \sum_{b=1}^{B} \sum_{s_1=1}^{M} \sum_{s_2=s_1+1}^{M} L_{ts} \tag{7}$$

where M represents the number of subsequences generated in each training step. During our training process, we set $M = 6$.

Self-dropout contrastive learning

The optimization goal of self-dropout was straightforward: the representations of a contig with different dropout masks were treated as positive pairs, while the representations of other contigs within the same batch of data were treated as negative pairs. The Transformer model was fed the identical contig twice, generating two representations of the contig with different dropout masks. During optimization, the representations of a contig with different dropout masks were pulled closer together, while the representations of other contigs within the batch were pushed far away. The loss function for self-dropout contrastive learning can be defined as:

$$\ell(\theta_{i,z_1}, \theta_{i,z_2}) = \frac{exp(cos(\theta_{i,z_1}, \theta_{i,z_2})/\tau)}{\sum_{b=1}^{B} 1_{b\neq i}[\sum_{j=1}^{2} exp(cos(\theta_{i,z_1}, \theta_{b,z_j})/\tau)] + exp(cos(\theta_{i,z_1}, \theta_{i,z_2})/\tau)} \tag{8}$$

$$L_{sd} = -log(\ell(\theta_{i,z_1}, \theta_{i,z_2})) \tag{9}$$

where θ_{i,z_1} and θ_{i,z_2} are the representations of the two representations of the i-th contig with z_1 and z_2 dropout masks.

The final loss function is: $L = \alpha L_{dmv} + \beta L_{sd}$. We set $\alpha = 2.0$ and $\beta = 1.0$ during our contrastive training.

Clustering

We employed a hierarchical two-stage clustering approach to cluster contigs using learned contig embeddings and single-copy gene (SCG) profiles in contigs. The methodology comprised three sequential phases: 1. initial clustering via the Leiden community detection algorithm [16] by contig embeddings, contig length, and SCG constraints; 2. contamination-based refinement of preliminary clusters using the FLSpp algorithm [17]; and 3. ensemble integration of clustering results obtained across multiple hyperparameter configurations by Galah for duplicate removal.

Initial clustering contigs with Leiden algorithm

Multiple k-nearest neighbor graphs of contigs were constructed using the same approach in COMEBin, with k parameters set to 75 and 100 connections. For each graph, its sparsification was applied at three edge retention levels (50%, 75%, and 100%) to evaluate clustering sensitivity across varying connectivity densities. The Leiden community detection algorithm was subsequently applied to each graph configuration with different hyperparameter settings. The resolution parameters in Leiden were set across seven discrete values (1, 2, 4, 6, 8, 10, 12), while all other Leiden parameters remained consistent with the COMEBin implementation. Phylogenetic coherence was maintained through constraint-based clustering, where contigs exhibiting identical SCG profiles were constrained to separate cluster assignments, thereby preserving taxonomic consistency within individual genome bins.

Refine MAGs with high contamination levels

Cluster quality evaluation followed the established approach in COMEBin. Contamination refinement was implemented according to the following criteria: Clusters exceeding 10% contamination underwent re-clustering via the FLSpp algorithm to enhance purity; clusters with moderate contamination ($5\% \leq$ contamination $\leq 10\%$) were subjected to removal of shorter contigs containing redundant SCGs; clusters with minimal contamination ($\leq 5\%$) remained unmodified. This iterative refinement process continued until contamination levels decreased below 5% or a maximum of three iterations was completed.

Ensemble MAGs from various hyperparameter configurations

Multiple clustering results were systematically selected according to three hyperparameters: resolution parameters in the Leiden algorithm, the k values, and edge retention levels of the k-nearest neighbor graph. For each parameter combination, we retained two optimal clustering results based on two selection criteria. The first selection prioritized maximizing the number of MAGs in this clustering result, which met quality thresholds ($\geq$ 90% completeness & $\leq$ 5% contamination). The secondary selection maximized MAGs count under relaxed criteria ($\geq$ 50% completeness & $\leq$ 10% contamination). Galah was subsequently applied to identify and remove redundant MAGs across all selected clustering results, ensuring retention of unique, high-quality MAGs.

Single-sample and Multi-sample Binning

For single-sample binning, metagenomic sequencing reads were subjected to independent assembly to generate sample-specific contigs. Read alignment was performed against the assembled contigs to compute coverage profiles for each contig. The resulting contigs and their corresponding coverage information served as input for the binning procedure. For multi-sample binning, metagenomic reads from each sample underwent independent assembly to produce sample-specific contigs. Subsequently, reads from all samples were mapped against each sample-specific contig to generate differential coverage profiles across samples. The sample-specific contigs and their cross-sample coverage matrices were then employed as input for multi-sample binning.

Benchmarking datasets

We evaluated CompleteBin performance against state-of-the-art binning algorithms using three CAMI II and five real-world datasets.

CAMI II datasets

CAMI II datasets were derived from the CAMI II challenge, encompassing three distinct environments: human microbiome from five body sites (Airways, 10 samples; Oral, 10 samples; Urogenital, 9 samples; Skin, 10 samples; Gastrointestinal, 10 samples), marine environments (10 samples), and plant-associated microbiomes (21 samples). Raw simulated sequencing reads were obtained from the organizers of the CAMI II challenge (https://data.cami-challenge.org) and assembled using MegaHit (v1.2.9) [37] with default parameters. Read alignment was performed using bwa-mem (v0.7.17) [38], with resulting BAM files sorted via samtools (v1.21) [39]. Contig coverage depth was computed using PySAM (v0.22.1). Contigs exceeding 900 base pairs were retained for CompleteBin analysis.

Real-world datasets

Real-world datasets comprised complex environmental samples from five ecosystems: marine (3 samples), freshwater (3 samples), plant-associated (1 sample), human fecal (3 samples), and cold seep sediments (6 samples). Assembly was conducted using MetaSPAdes (v4.2.0) [40] with default parameters, followed by identical read processing procedures as described for CAMI II datasets. Contig length thresholds were set at 900 base pairs for all datasets except cold seep sediments, where a threshold of 768 base pairs was applied to accommodate the super complexity of this environment.

Evaluation metrics

MAG quality assessment was performed using CheckM2 (v1.0.1) to determine completeness and contamination levels. MAGs were categorized according to established quality thresholds: near-complete MAGs (completeness $\geq$90% and contamination $\leq$5%), medium-quality MAGs (completeness $\geq$50% and contamination $\leq$10%, excluding those meeting near-complete criteria), and low-quality MAGs (all remaining bins not satisfying the aforementioned criteria). Binning algorithm performance was quantified by counting near-complete and medium-quality MAGs across both CAMI II and real-world datasets.

For CAMI II datasets, the F1 score (bp) calculation was performed through a two-step process. Initially, contig-to-reference genome alignment was conducted using Minimap2 (v2.28) with default parameters. Subsequently, F1 scores (bp) were computed following the VAMB methodology, which evaluates performance based on correctly aligned nucleotide positions. For real-world datasets, the mapping ratio was calculated using the BAM files of each real-world metagenomic sequencing sample.

Experimental configuration for binning algorithms

We benchmarked CompleteBin against five established binning algorithms: VAMB (v3.0.2), CONCOCT (v1.1.0), MetaBAT2 (v2.15), SemiBin2 (v2.1.0), and COMEBin (v1.0.4). Each algorithm was configured with its minimum acceptable contig length threshold: VAMB (>2,000 bp), CONCOCT (>1,000 bp), MetaBAT2 (>1,500 bp), SemiBin2 (>1,000 bp), and COMEBin (>1,000 bp). Algorithm-specific parameters were set as follows: CONCOCT with 200 iterations, SemiBin2 without reclustering, and MetaBAT2, VAMB, and COMEBin with default configurations. CompleteBin was implemented in Python with neural network components developed using PyTorch (v2.1.0). The SCGs' calling was executed using Prodigal (v2.6.3) and HMMER (v3.3.2). The model would be trained for 36 epochs with a batch size of 1,024. All experiments were conducted on a high-performance computing node equipped with dual AMD EPYC 7742 processors (64 cores, 128 threads) and 1 TB RAM. Neural

network training was accelerated using a single NVIDIA Tesla A100-40GB GPU.

Microbial diversity estimation and MAG-to-reference mapping

Microbial diversity within metagenomic samples was quantified using Nonpareil [41] (v3.303) with parameters '-T alignment -f fasta' based on read redundancy analysis. For each MAG, the reference genome exhibiting the highest F1 score (bp) was designated as the corresponding reference genome of that MAG.

Metagenomic analysis of cold seep sediment samples

Medium-quality and near-complete MAGs obtained from all binning tools were pooled together and then clustered at a 0.99 similarity threshold using dRep [42] (v3.5.0) to generate StGBs. Representative MAGs of these StGBs were taxonomically annotated using GTDB-Tk [43] (v2.4.0). Taxonomic annotations of other MAGs within the same clustering group were inherited from their representative MAG. After filtering out all medium-quality MAGs from the clustering result, StGBs that could still be reported by a specific tool were considered detectable by that tool.

Gene prediction was conducted on all NCMAGs obtained by any of the binning tools using Prodigal [44] (v2.6.3). Predicted genes were clustered using CD-HIT [45] (v4.8.1) at a 0.95 similarity threshold. Representative genes that were obtained from NCMAGs that were exclusively recovered by CompleteBin were counted and annotated using eggNOG-mapper [46] (v2.1.12).

For each sample, genome-scale metabolic models were constructed for all NCMAGs recovered from it using CarveMe [47] (v1.6.0), which were then used to build community-level metabolic models for each sample and each binning tool. For each community-level model, a specific substance was counted only once in statistics. Sub-models of nitrogen and sulfur metabolism were respectively manually extracted from each community-level metabolic network in Cytoscape [48] (v3.10.2) to generate the community-level sub-models. For each reaction step, all StGBs involved in that step were identified. By integrating all community-level sub-models obtained from all six samples, we identified StGBs capable of performing each step that were detected by at least one binning tool. Among these, StGBs exclusively detected by CompleteBin were subjected to statistical analysis and visualization.

Statistics

Statistical analyses were performed using R (v4.1.1). Pairwise comparisons of paired datasets were evaluated using Wilcoxon signed-rank tests, while independent group comparisons employed Wilcoxon rank-sum tests, both implemented via the 'wilcox.test' function. Linear regression modeling was conducted using the 'lm' function to assess relationships between continuous variables.

Data availability

CAMI II datasets were obtained from the CAMI II challenge repository (https://cami-challenge.org/datasets/). Real-world metagenomic sequencing data were sourced from multiple public repositories. Freshwater samples were retrieved from the NCBI Sequence Read Archive (SRA) under accessions SRR26420192, ERR9631077, and ERR4195020. The plant-associated sample was accessed via SRA accession SRR10968246, while human fecal samples were obtained under accessions SRR13060973 and SRR13060977. Marine samples were acquired from the Joint Genome Institute (JGI) database under project identifiers 1021526, 1125694, and 1102218, as described in [49]. Cold seep sediment samples were sourced from [25] and downloaded from the China National GeneBank (CNGB) database under project accession CNP0005072, with individual run identifiers CNR1069767 (CS01), CNR1069770 (CS03), CNR1069773 (CS05), CNR1069776 (CS08), CNR1069779 (CS10), and CNR1069782 (CS12). An additional human fecal metagenomic sample from the IBS-D cohort was obtained from CNGB under project accession CNP0000334 (run ID: CNR0055062, sample K0268).

Code availability

The source code is freely available at https://github.com/ericcombiolab/CompeleBin [50] under an MIT license.

Acknowledgements

The project is supported by the Young Collaborative Research grant (No. C2004-23Y, No. CRF-C2005-24Y), HMRF grant (No. 11221026), and HKBU RCMS (RCMS/24-25/03).

Competing Interests

The authors declare no competing interests.

Figure Legends/Captions

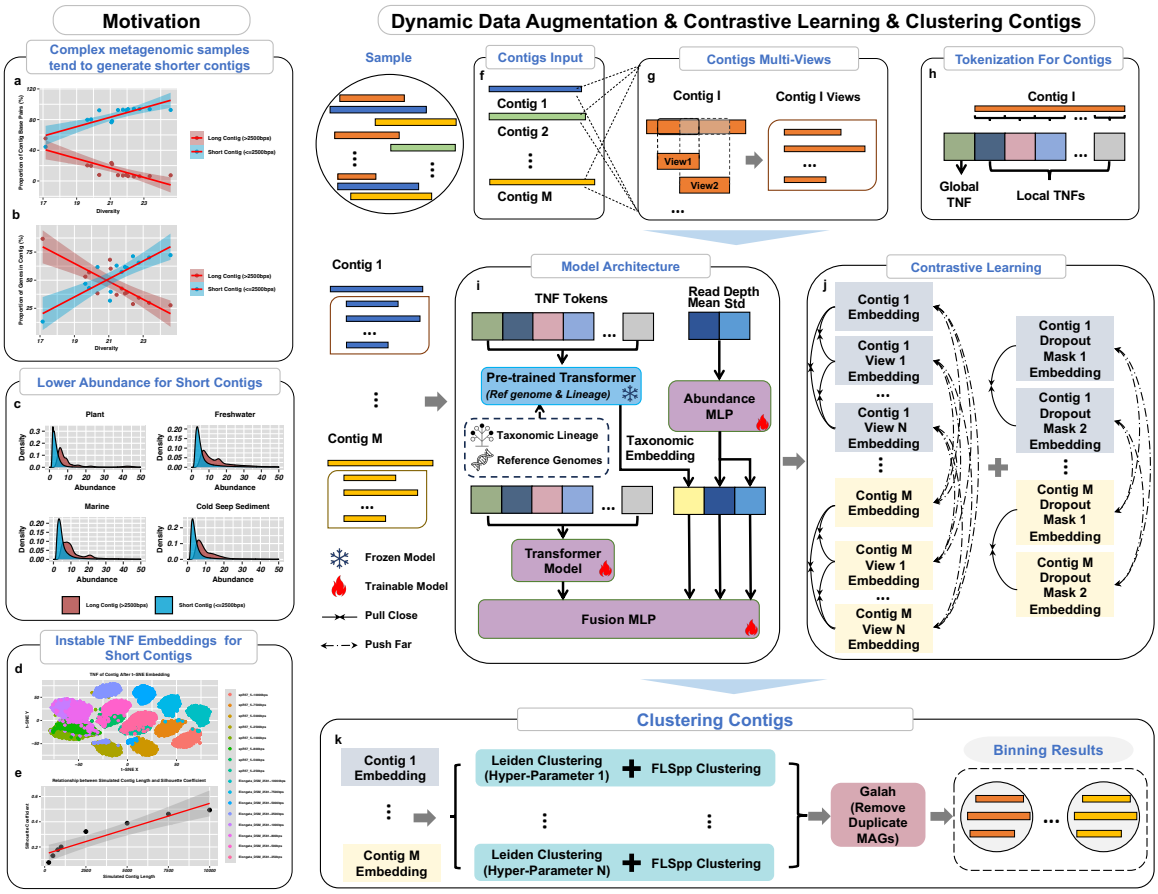

Figure 1: Motivation and framework of CompleteBin. (**a**). The contribution of short contigs to contig base pairs. (**b**). The contribution of short contigs to complete protein-coding genes. (**c**). Abundance distributions of short and long contigs in Plant, Freshwater, Marine, and Cold Seep Sediment environments. (**d**) t-SNE visualization of TNF features across contigs of different lengths. Each point represents the t-SNE embedding of TNF features from each contig, with colors indicating different simulation conditions: contigs were generated by randomly extracting sequences from the spR57 and Elongata_DSM_2581 reference genomes, with lengths ranging from 250 bp to 10,000 bp. (**e**). Silhouette coefficient analysis for the simulated sequences from spR57 and Elongata_DSM_2581. (**f**). CompleteBin processes a batch with M contigs, and dynamic data augmentation is performed on these contigs at each training step. (**g**). A dynamic data augmentation strategy generates multiple views (subsequences) from each contig within training batches. (**h**). Context-based TNF tokenization is applied for each contig and its augmented views. (**i**). CompletBin integrates sequence and abundance information through three parallel pathways: (1) a frozen pre-trained Transformer processing context-based TNF tokens to generate static taxonomic embeddings of contigs, (2) a trainable Transformer processing identical tokens for dynamic representation of contigs, and (3) an MLP encoding read depth statistics (mean, standard deviation). Outputs are concatenated and passed through a fusion MLP to generate final contig embeddings. (**j**). The training objective minimizes distances between (1) a contig and its augmented views (positive pairs) while maximizing distances from other contigs (negative pairs), and (2) embeddings of the same contig generated with different dropout masks (positive pairs) while maximizing distances from other contigs (negative pairs). (**k**). Contigs are clustered using Leiden and FLSpp algorithms based on their embeddings and single-copy gene profiles across multiple hyperparameter configurations. The Galah tool integrates these clustering results to generate the final bins.

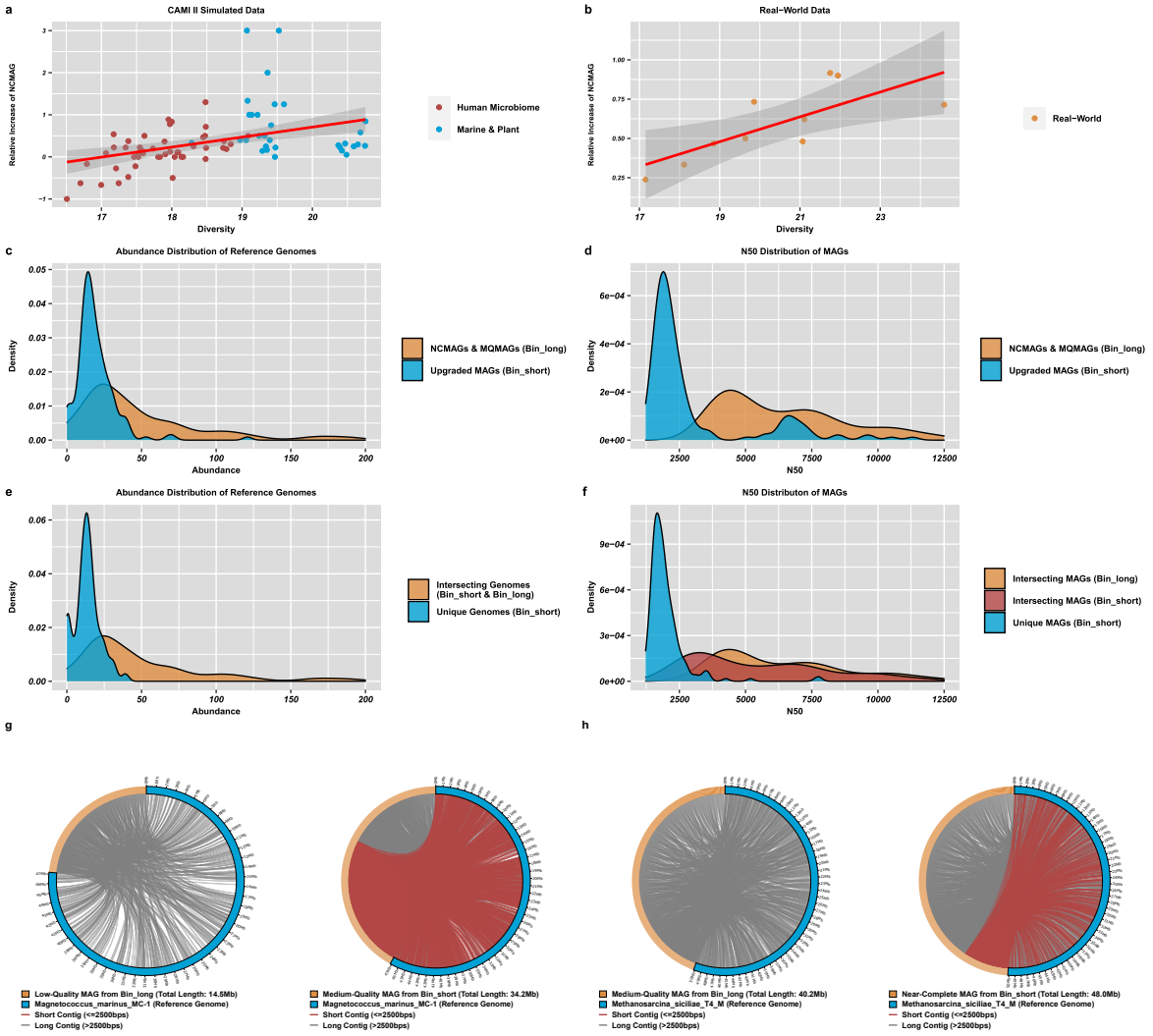

Figure 2: (**a**). Correlation between microbial diversity and the relative increase of NCMAGs between CompleteBin and COMEBin with CAMI II datasets. (**b**). Correlation between microbial diversity and the relative increase of NCMAG between CompleteBin and COMEBin with real-world metagenomic samples. (**c**). Abundance distributions of reference genomes corresponding to NCMAGs and MQMAGs recovered in Bin_long, compared with upgraded MAGs in Bin_short. (**d**). Assembly contiguity (N50) distributions for NCMAGs and MQMAGs in Bin_long, compared with upgraded MAGs in Bin_short. (**e**). Abundance distributions of reference genomes for MAGs recovered in both Bin_long and Bin_short (intersecting genomes) versus genomes unique to Bin_short. (**f**). Assembly contiguity (N50) distributions for MAGs of intersecting genomes (present in both Bin_long and Bin_short) versus MAGs unique to Bin_short. (**g**). A low-quality MAG from Bin_long (reference: *Magnetococcus marinus MC-1*) upgraded to MQMAG in Bin_short through the incorporation of short contigs. (**h**). An MQMAG from Bin_long (reference: *Methanosarcina siciliae T4/M*) upgraded to NCMAG in Bin_short through the incorporation of short contigs.

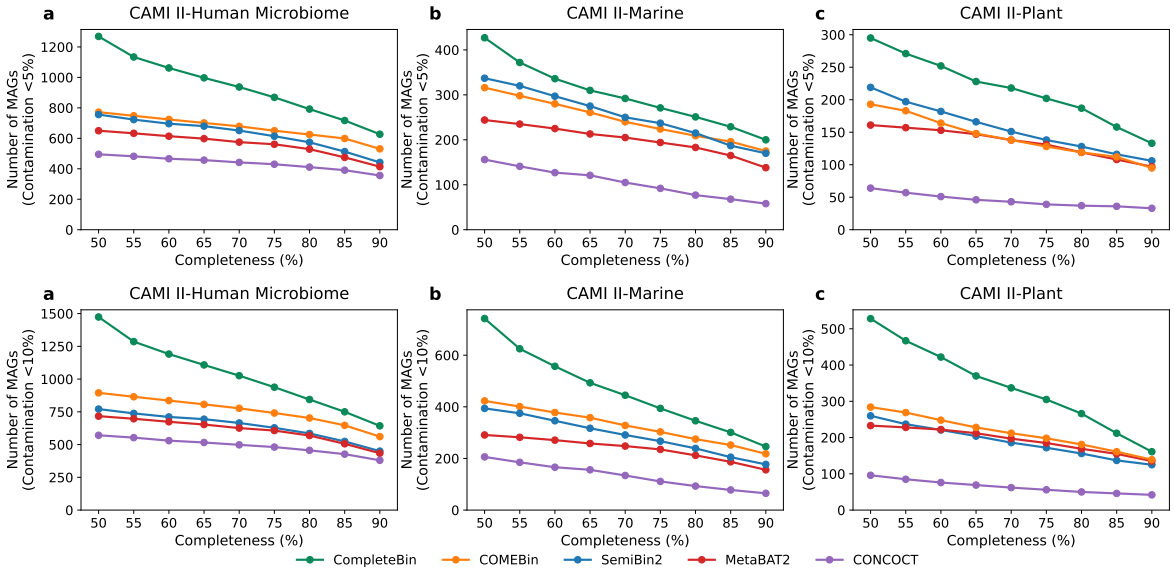

Figure 3: Comparison of binning methods on the CAMI II human microbiome (**a**, **d**), marine (**b**, **e**), and plant (**c**, **f**) datasets. The number of recovered MAGs with contamination thresholds of ≤5% and ≤10% is shown with different completeness thresholds.

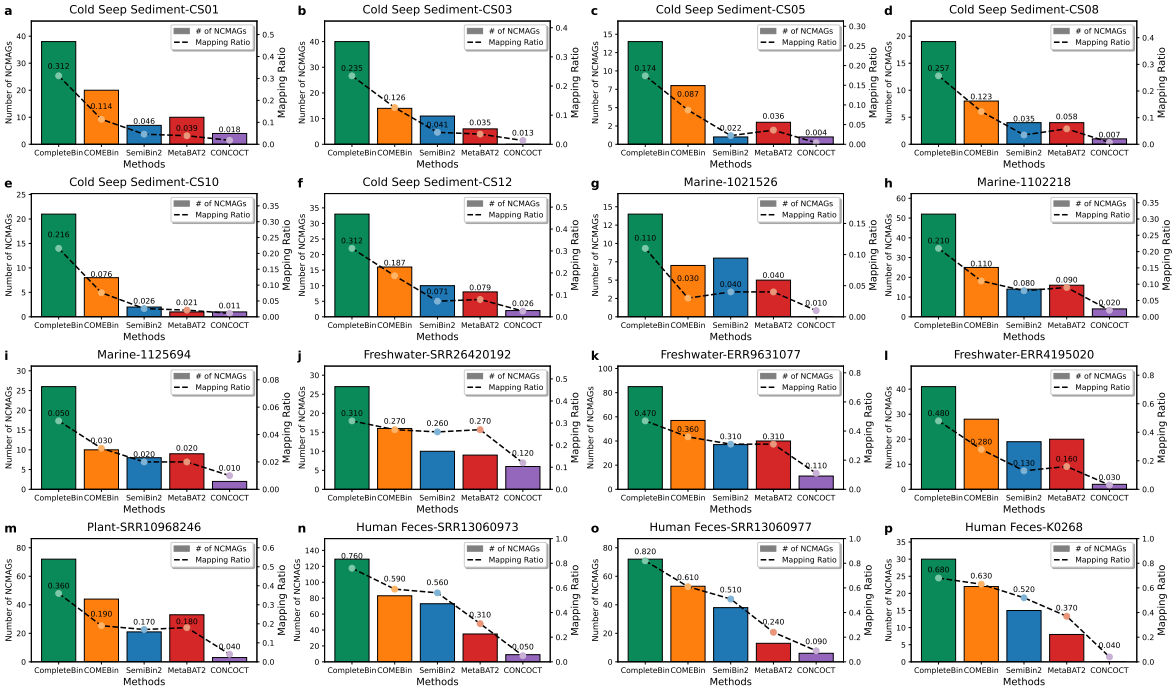

Figure 4: The number of recovered NCMAGs and the mapping ratios of different binning methods on the cold seep sediments (**a-f**), marine (**g-i**), freshwater (**j-l**), plant-associated (**m**), and human fecal (**n-p**) metagenomic sequencing samples.

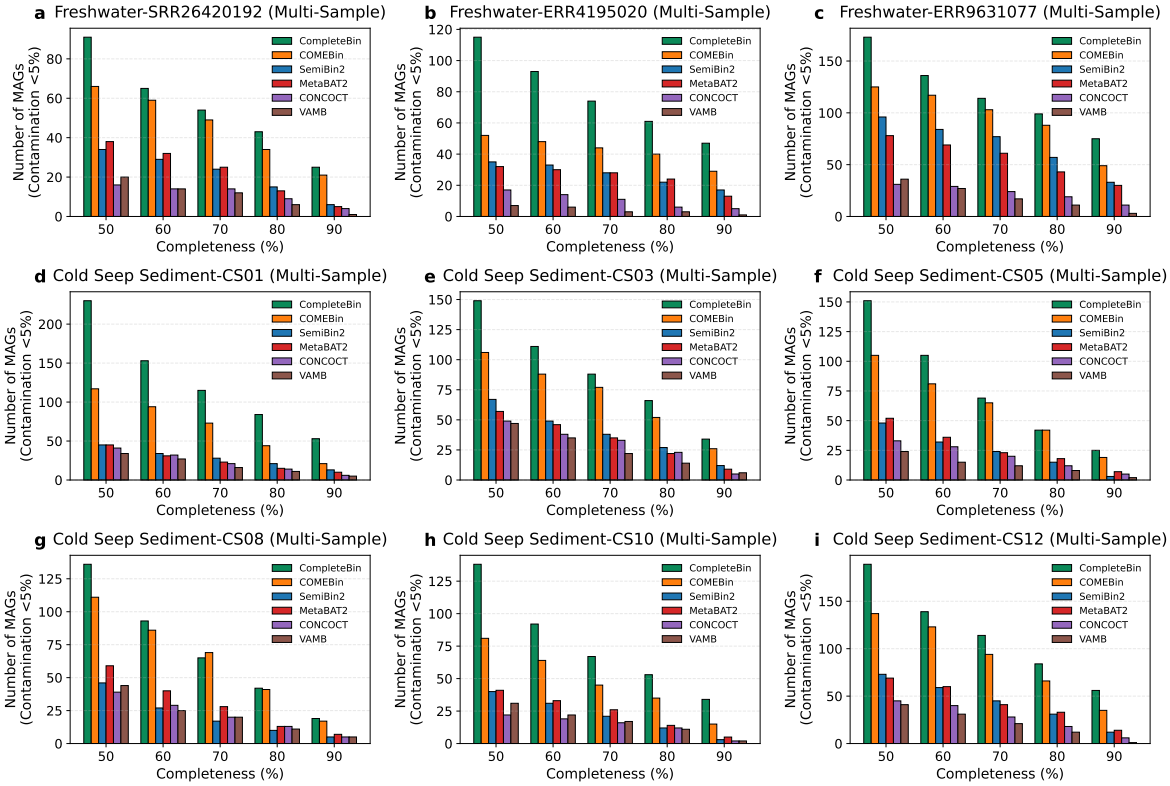

Figure 5: Comparison of binning methods on the three freshwater (**a-c**), and six cold seep sediment (**d-i**) sequencing samples with multi-sample binning mode. The number of recovered MAGs with contamination thresholds of $\leq$5% is shown for varying completeness thresholds.

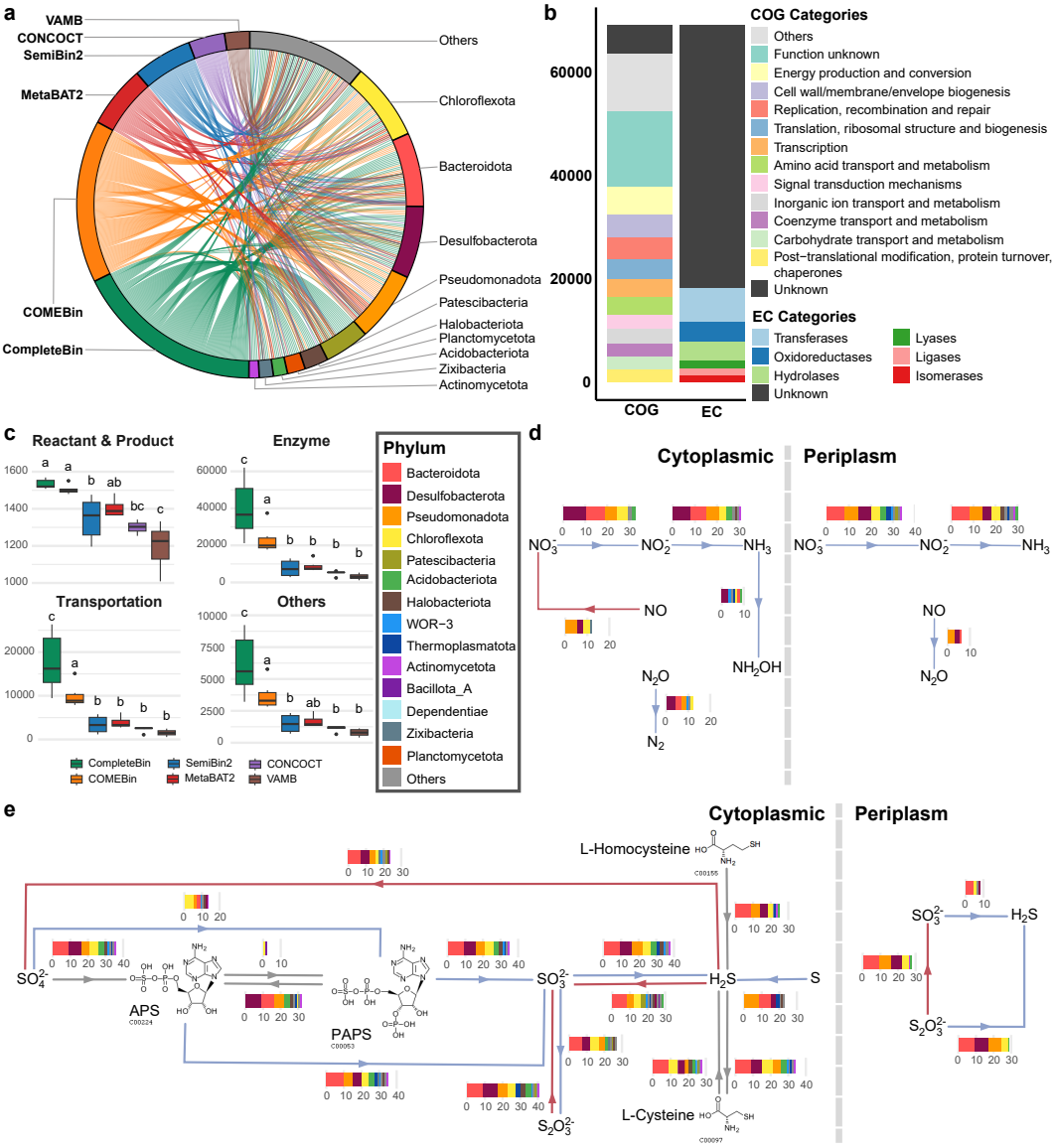

Figure 6: (**a**). Near-complete MAGs were detected in cold seep sediments. The left semicircle shows 6 different binners, while the right semicircle displays different phyla. Each connecting line between semicircles represents the left connected binner recovered a StGB from the right connected Phyla. (**b**). Annotation of complete genes assigned to MAGs by CompleteBin but not by other tools. The left and right parts are classified by COG and EC. (**c**). Statistical summary of basic information regarding the metabolic network. The four graphs respectively represent the distribution of reactants and products, enzymatic reactions, transport reactions, and other reactions in the network. Compact letter displays denote statistically significant differences among groups (one-way ANOVA followed by Tukey's honest significant difference post hoc test, $\alpha = 0.05$). (**d**) and (**e**). Metabolic networks of nitrogen (**d**) and sulfur (**e**). Blue lines indicate reduction reactions, and red lines represent oxidation reactions. Stacked bar charts show the phylum-level distribution of StGBs uniquely identified by CompleteBin that are involved in the reactions of corresponding steps.

References

[1] Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 1–22 (2020).

[2] Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics* **22**, 178–193 (2021).

[3] Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology* **35**, 833–844 (2017).

[4] Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal* **19**, 6301–6314 (2021).

[5] Wang, J. & Jia, H. Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology* **14**, 508–522 (2016).

[6] Wang, Z. *et al.* Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nature Communications* **15**, 585 (2024).

[7] Johansen, P. L., Chatzigiannidou, I., Berzina, L., Kristiansen, K. & Brix, S. Unveiling soil microbial diversity through ultra-deep short-read metagenomic sequencing and co-assembly. *Imeta* e70075 (2025).

[8] Yorki, S. *et al.* Comparison of long-and short-read metagenomic assembly for low-abundance species and resistance genes. *Briefings in Bioinformatics* **24**, bbad050 (2023).

[9] Kang, D. D. *et al.* Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

[10] Liu, C.-C. *et al.* Metadecoder: a novel method for clustering metagenomic contigs. *Microbiome* **10**, 46 (2022).

[11] Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nature biotechnology* **39**, 555–560 (2021).

[12] Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 1–18 (2014).

[13] Alneberg, J. *et al.* Concoct: clustering contigs on coverage and composition. *arXiv preprint arXiv:1312.4038* (2013).

[14] Pan, S., Zhao, X.-M. & Coelho, L. P. Semibin2: self-supervised contrastive learning leads to better mags for short-and long-read sequencing. *Bioinformatics* **39**, i21–i29 (2023).

[15] Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nature Methods* **20**, 1170–1173 (2023).

[16] Traag, V. A., Waltman, L. & Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports* **9**, 1–12 (2019).

[17] Conrads, T., Drexler, L., Könen, J., Schmidt, D. R. & Schmidt, M. Local search k-means++ with foresight. *arXiv preprint arXiv:2406.02739* (2024).

[18] Meyer, F. *et al.* Critical assessment of metagenome interpretation: the second round of challenges. *Nature methods* **19**, 429–440 (2022).

[19] Han, H., Wang, Z. & Zhu, S. Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes. *Nature Communications* **16**, 2865 (2025).

[20] Duncan, A. *et al.* Metagenome-assembled genomes of phytoplankton microbiomes from the arctic and atlantic oceans. *Microbiome* **10**, 67 (2022).

[21] Buck, M. *et al.* Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. *Scientific Data* **8**, 131 (2021).

[22] Kavagutti, V. S. *et al.* High-resolution metagenomic reconstruction of the freshwater spring bloom. *Microbiome* **11**, 15 (2023).

[23] Maestre-Carballa, L., Navarro-López, V. & Martinez-Garcia, M. City-scale monitoring of antibiotic resistance genes by digital pcr and metagenomics. *Environmental Microbiome* **19**, 16 (2024).

[24] Tláskal, V. *et al.* Metagenomes, metatranscriptomes and microbiomes of naturally decomposing deadwood. *Scientific data* **8**, 198 (2021).

[25] Wang, X. *et al.* A scalable practice for deep-sea metagenomic studies. *bioarxiv* (2024).

[26] Kim, C. Y. *et al.* Human reference gut microbiome catalog including newly assembled genomes from under-represented asian metagenomes. *Genome Medicine* **13**, 1–20 (2021).

[27] Zhao, L. *et al.* A clostridia-rich microbiota enhances bile acid excretion in diarrhea-predominant irritable bowel syndrome. *The Journal of clinical investigation* **130**, 438–450 (2024).

[28] Burian, J. *et al.* Bioactive molecules unearthed by terabase-scale long-read sequencing of a soil metagenome. *Nature Biotechnology* 1–8 (2025).

[29] Kim, C., Pongpanich, M. & Porntaveetus, T. Unraveling metagenomics through long-read sequencing: a comprehensive review. *Journal of Translational Medicine* **22**, 111 (2024).

[30] Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[31] Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks* **107**, 3–11 (2018).

[32] Zhang, B. & Sennrich, R. Root mean square layer normalization. *Advances in neural information processing systems* **32** (2019).

[33] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929–1958 (2014).

[34] Agarap, A. F. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[35] Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456 (pmlr, 2015).

[36] Mende, D. R. *et al.* progenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic acids research* gkw989 (2016).

[37] Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

[38] Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997* (2013).

[39] Li, H. *et al.* The sequence alignment/map format and samtools. *bioinformatics* **25**, 2078–2079 (2009).

[40] Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaspades: a new versatile metagenomic assembler. *Genome research* **27**, 824–834 (2017).

[41] Rodriguez-r, L. M. & Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**, 629–635 (2014).

[42] Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. drep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal* **11**, 2864–2868 (2017).

[43] Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database (2020).

[44] Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**, 1–11 (2010).

[45] Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

[46] Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggnog-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* **38**, 5825–5829 (2021).

[47] Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research* **46**, 7542–7553 (2018).

[48] Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).

[49] Duncan, A. *et al.* Metagenome-assembled genomes of phytoplankton microbiomes from the arctic and atlantic oceans. *Microbiome* **10**, 67 (2022).

[50] Zou, B. The version 1.0.7 of completebin. *Zenodo* (2025). URL `https://doi.org/10.5281/zenodo.17451857`.