nature communications

Article

https://doi.org/10.1038/s41467-025-57957-6

Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes

Received: 5 March 2024

Accepted: 7 March 2025

Published online: 24 March 2025

Check for updates

Haitao Han[1,6], Ziye Wang[2,6] & Shanfeng Zhu[1,3,4,5]

Metagenomic binning is a culture-free approach that facilitates the recovery of metagenome-assembled genomes by grouping genomic fragments. However, there remains a lack of a comprehensive benchmark to evaluate the performance of metagenomic binning tools across various combinations of data types and binning modes. In this study, we benchmark 13 metagenomic binning tools using short-read, long-read, and hybrid data under co-assembly, single-sample, and multi-sample binning, respectively. The benchmark results demonstrate that multi-sample binning exhibits optimal performance across short-read, long-read, and hybrid data. Moreover, multi-sample binning outperforms other binning modes in identifying potential antibiotic resistance gene hosts and near-complete strains containing potential biosynthetic gene clusters across diverse data types. This study also recommends three efficient binners across all data-binning combinations, as well as high-performance binners for each combination.

Microorganisms serve as the engines driving Earth's biogeochemical cycles[1]. Nevertheless, the majority of microbial diversity remains undiscovered, with estimates indicating that only a small fraction of microbial species have been identified and characterized[2]. Furthermore, many microbes are difficult to study in isolation as they are not yet cultivatable in laboratory[3]. De novo assembly of DNA sequences from metagenomes and binning into metagenome-assembled genomes (MAGs) has emerged as an alternative strategy to explore unknown microbial communities[4,5]. These MAGs substantially expand the microbial tree of life and offer insights into microbial ecological characteristics[6,7].

Over the past decade, several tools have been developed for metagenomic binning, involving the assignment of assembled contigs to MAGs based on features of sequence composition and coverage profiles (Table 1). For instance, CONCOCT[8] integrates sequence composition and coverage as contig features, performs dimensionality reduction using principal component analysis (PCA), and ultimately utilizes a Gaussian mixture model (GMM) for contig clustering. MaxBin 2[9] estimates the likelihood of a given contig belonging to a particular genome by utilizing tetranucleotide frequencies and contig coverages. This probability is then used in an Expectation-Maximization (EM) algorithm to recover genomes from metagenomes. MetaBAT 2[10] calculates pairwise similarities between contigs using tetranucleotide frequency and contig coverage, and it utilizes the resulting similarity graph to perform contig clustering via a modified label propagation algorithm (LPA)[11]. More recently, Binny[12] applies multiple k-mer compositions and contig coverage for iterative, non-linear dimensionality reduction, followed by an iterative clustering approach employing hierarchical density-based spatial clustering of applications with noise (HDBSCAN)[13]. MetaDecoder[14] uses a modified Dirichlet process Gaussian mixture model (DPGMM) for preliminary clustering, followed by a semi-supervised k-mer frequency probabilistic model and a modified GMM for further clustering. MetaBinner[15] is a stand-alone ensemble algorithm that employs "partial seed" k-means and multiple types of

[1]Institute of Science and Technology for Brain-Inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai, China. [2]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China. [3]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Ministry of Education, Shanghai, China. [4]Shanghai Key Lab of Intelligent Information Processing and Shanghai Institute of Artificial Intelligence Algorithm, Fudan University, Shanghai, China. [5]Zhangjiang Fudan International Innovation Center, Shanghai, China. [6]These authors contributed equally: Haitao Han, Ziye Wang. e-mail: zhusf@fudan.edu.cn

Table 1 | Overview of benchmarked metagenomic binning tools

Tool	Description	Year	Reference
Stand-alone binner			
CONCOCT	Uses PCA for dimension reduction and GMM for clustering	2014	8
MaxBin 2	Uses EM algorithm to assign contigs to MAGs	2015	9
MetaBAT 2	Constructs a similarity graph and utilizes LPA for partitioning	2019	10
VAMB	Uses VAE to encode contigs	2021	16
CLMB	Uses data augmentation and contrastive learning to learn representation for the contigs	2022	17
MetaDecoder	Uses DPGMM for initial clustering and a semi-supervised probability model along with a modified GMM for subsequent clustering	2022	14
Binny	Uses PCA and t-SNE for dimension reduction and HDBSCAN for iterative clustering	2022	12
MetaBinner	Uses "partial seed" K-means to generate multiple binning results with different types of features and a two-stage ensemble strategy to integrate these results	2023	15
SemiBin 2	Constructs must-link and cannot-link constraints and combines them with contrastive learning to derive feature embeddings of the contigs.	2023	19
COMEBin	Uses data augmentation to generate multi-views for each contig and combines contrastive learning to obtain high-quality embeddings.	2024	20
Bin-refinement tool			
DAS Tool	Aggregates candidate bins, iteratively selects high-scoring bins, and updates the remaining bins	2018	23
MetaWRAP	Uses binning_refiner[68] to produce hybrid bin sets, selects better bins from similar bins, and removes duplicate contigs	2018	22
MAGScoT	Creates hybrid bins and then performs iterative scoring and refinement	2022	24

features to generate component results. It then utilizes a two-stage ensemble strategy to integrate these component results.

Recently, a series of deep learning-based binning methods have been proposed. VAMB[16] uses deep variational autoencoders (VAE) to encode tetranucleotide frequency and coverage information for each contig, the latent representation is then processed using an iterative medoid clustering algorithm. Subsequently, CLMB[17] introduces simulated noise for each contig and utilizes contrastive learning to produce more robust contig embeddings. SemiBin 1[18] is a semi-supervised binning algorithm that relies on deep siamese neural networks to effectively leverage must-link and cannot-link information. SemiBin 2[19] improves upon SemiBin 1 by utilizing self-supervised learning to learn feature embeddings from the contigs. Furthermore, it introduces a novel ensemble-based DBSCAN approach designed specifically for long-read data. More recently, COMEBin[20] introduces data augmentation to generate multiple views for each contig, combines them with contrastive learning to obtain high-quality (HQ) embeddings, and then applies a Leiden-based[21] method for clustering. Additionally, MetaWRAP[22], DAS Tool[23], and MAGScoT[24] are bin-refinement tools that combine the strengths of multiple binning tools to reconstruct the HQ MAGs.

Metagenomic binning comprises three modes: co-assembly, single-sample, and multi-sample binning. Co-assembly binning initially assembles all sequencing samples, and the resulting contigs are then binned with coverage information calculated across samples. This mode can leverage co-abundance information[18], but it may result in inter-sample chimeric contigs[25] and is unable to retain sample-specific variation[16]. Single-sample binning means assembling and binning independently within each sample. Multi-sample binning differs from single-sample binning in that it calculates coverage information across samples. Although this process is time-consuming, it often results in the recovery of higher-quality MAGs[26].

Over the past few years, several benchmarking studies have been introduced to facilitate a fair and comprehensive evaluation of metagenomic binning tools[26–30]. However, existing research has not thoroughly assessed the performance of metagenomic binning tools when considering different combinations of data types and binning modes (data-binning combinations). Each data-binning combination denotes the utilization of a specific binning mode with a particular data type (see Table 2 and Supplementary Fig. 1b). Moreover, previous benchmarking studies have not taken into account the continuous development of new binning algorithms and tools for genome quality assessment, such as CheckM 2[31].

In this work, we evaluate 13 metagenomic binning tools (Table 1) using seven data-binning combinations on five real-world datasets (Supplementary Tables 1–5) with metagenomic next-generation sequencing (mNGS), PacBio high-fidelity (HiFi) and Oxford Nanopore data (Supplementary Fig. 1). The number of recovered "moderate or higher" quality (MQ, see Section Evaluation metrics and ranking score), near-complete (NC), and HQ MAGs assessed by CheckM 2 illustrates that multi-sample binning achieves an average improvement of 125%, 54%, and 61% compared to single-sample binning on marine short-read, long-read, and hybrid data, respectively. Subsequently, we identify the top three high-performance binners for each data-binning combination (Table 2). COMEBin and MetaBinner rank first in four and two data-binning combinations, respectively. Binny ranks first in the short_co data-binning combination. MetaBAT 2, VAMB, and MetaDecoder are highlighted as efficient binners due to their excellent scalability (see Table 2). Furthermore, MetaWRAP, DAS Tool, and MAGScoT are employed to refine the MAGs recovered by the top three binners. Among them, MetaWRAP demonstrates the best overall performance in recovering MQ, NC, and HQ MAGs, while MAGScoT achieves comparable performance and excellent scalability. The refined MAGs from MAGScoT are then utilized for subsequent analysis (Supplementary Fig. 1c). Additionally, MAG dereplication is performed to analyze the diversity of species and strains for seven data-binning combinations. Antibiotic Resistance Genes (ARGs) and Biosynthetic Gene Clusters (BGCs) are then annotated in the refined non-redundant MAGs. Multi-sample binning demonstrates remarkable superiority over single-sample binning by identifying 30%, 22%, and 25% more potential ARG hosts, as well as 54%, 24%, and 26% more potential BGCs from NC strains, across short-read, long-read, and hybrid data, respectively. Based on the results of dereplication and annotation, we provide the recommended usage order for seven data-binning combinations (Table 2).

Results

Benchmarking binners on seven data-binning combinations

We evaluated the performance of ten stand-alone binners on real datasets under seven data-binning combinations (Fig. 1 and

Table 2 | Overview of data-binning combinations and binner performance: detailed configurations of seven combinations, the top three high-performance binners for each combination, and three overall efficient binners

Data-binning combination	Data	Assembled sample	Binning mode
Hybrid_multi	Short and long read	Single sample	Multi-sample
Hybrid_single	Short and long read	Single sample	Single-sample
Long_multi	Long read	Single sample	Multi-sample
Long_single	Long read	Single sample	Single-sample
Short_multi	Short read	Single sample	Multi-sample
Short_single	Short read	Single sample	Single-sample
Short_co	Short read	All samples	Co-assembly
Data-binning combination	**Top three high-performance binners**		
Hybrid_multi	COMEBin	Binny	MetaBinner
Hybrid_single	COMEBin	MetaDecoder	SemiBin 2
Long_multi	MetaBinner	COMEBin	SemiBin 2
Long_single	MetaBinner	SemiBin 2	MetaDecoder
Short_multi	COMEBin	Binny	MetaBinner
Short_single	COMEBin	MetaDecoder	SemiBin 2
Short_co	Binny	SemiBin 2	MetaBinner
Efficient binners			
MetaBAT 2, VAMB, MetaDecoder			

The data-binning combinations were listed in descending order based on their overall ability to recover HQ MAGs. The high-performance binners were listed according to their overall ranking scores (see Section Evaluation metrics and ranking score), while the efficient binners were listed based on the computational resources they used. The full rankings of high-performance binners are shown in Fig. 2.

Supplementary Fig. 2). Following the guidelines of the second CAMI challenge (CAMI II)[28] and the Minimum Information about a Metagenome-Assembled Genome[32], we denoted MAGs with completeness >50% and contamination <10% as "moderate or higher" quality (MQ) MAGs. Those with completeness >90% and contamination <5% were considered as NC MAGs. HQ MAGs were defined as those with completeness >90%, contamination <5%, and the presence of 23S, 16S, and 5S rRNA genes, as well as at least 18 tRNAs. The median number of recovered MAGs across the ten binners for each data-binning combination was calculated for subsequent analysis.

For short-read data, co-assembly binning recovered the fewest number of MQ, NC, and HQ MAGs across five datasets (Fig. 1 and Supplementary Fig. 2). In the human gut I dataset (three mNGS samples), multi-sample binning recovered 22% (181 versus 148) more MQ MAGs and retrieved comparable numbers of NC and HQ MAGs compared to single-sample binning. In the human gut II dataset (30 mNGS samples), multi-sample binning surpassed single-sample binning by recovering 44% (1908 versus 1328) more MQ MAGs, 82% (968 versus 531) more NC MAGs, and 233% (100 versus 30) more HQ MAGs. In the marine dataset (30 mNGS samples), multi-sample binning substantially outperformed single-sample binning, retrieving 100% (1101 versus 550) more MQ MAGs, 194% (306 versus 104) more NC MAGs, and 82% (62 versus 34) more HQ MAGs. Moreover, multi-sample binning demonstrated better performance compared to single-sample binning in both the cheese dataset (15 mNGS samples) and the activated sludge dataset (23 mNGS samples).

For long-read data (Fig. 1 and Supplementary Fig. 2), the performance of the multi-sample and single-sample binning is comparable in the human gut I (three PacBio HiFi samples) and cheese (15 PacBio HiFi samples) datasets. However, in the marine dataset (30 PacBio HiFi samples), the multi-sample binning exhibited substantial improvements, recovering 50% (1196 versus 796) more MQ, 55% more (191 versus 123) NC, and 57% more (163 versus 104) HQ MAGs compared to the single-sample binning. Additionally, multi-sample binning exhibited better results compared to single-sample binning in the human gut II dataset (30 Nanopore samples) and the activated sludge dataset (23 Nanopore samples). The results demonstrated that multi-sample binning of long-read data required a larger number of samples than short-read data to demonstrate substantial improvements. This difference may be due to the relatively lower sequencing depth in third-generation sequencing compared to mNGS in these datasets[29,33,34].

For hybrid data (Fig. 1 and Supplementary Fig. 2), multi-sample binning slightly outperformed single-sample binning in recovering more MQ, NC and HQ MAGs from the human gut I dataset (three hybrid samples). In the marine dataset (30 hybrid samples), multi-sample binning demonstrated remarkable improvement, retrieving 46% (1650 versus 1131) more MQ MAGs, 70% (473 versus 278) more NC MAGs, and 66% (332 versus 200) HQ MAGs compared to single-sample binning. Similarly, multi-sample binning achieved better results over single-sample binning in both the cheese dataset (15 mNGS samples) and the human gut II dataset (30 mNGS samples). In the activated sludge dataset, CONCOCT, MaxBin 2, Binny, MetaBinner, and COMEBin consistently demonstrated excellent performance in the multi-sample binning mode compared to the single-sample binning mode.

Overall, multi-sample binning consistently outperformed single-sample binning across short-read, long-read and hybrid data. Additionally, short-read data exhibited poor performance in recovering HQ MAGs. This suggested that the continuity advantages of long-read and hybrid data led to the recovery of more HQ MAGs. Moreover, hybrid data, which combined the strengths of short-read and long-read data[35], demonstrated the best overall performance in both the single-sample and multi-sample binning across five datasets (Fig. 1b, d, f, h, and j). The recommended usage order for seven data-binning combinations is listed in Table 2.

Determining the high-performance and efficient binners

To evaluate the performance of the ten stand-alone binners, we calculated overall ranking scores (see Section Evaluation metrics and ranking score) across seven data-binning combinations based on the number of MQ, NC, and HQ MAGs they recovered from five datasets (Fig. 2a). Detailed ranking scores for each dataset were shown in Supplementary Fig. 3. Notably, CONCOCT, MaxBin 2, MetaBinner, and COMEBin (GPU) did not complete execution within two weeks on the short-read co-assembly data from the activated sludge dataset with

https://doi.org/10.1038/s41467-025-57957-6

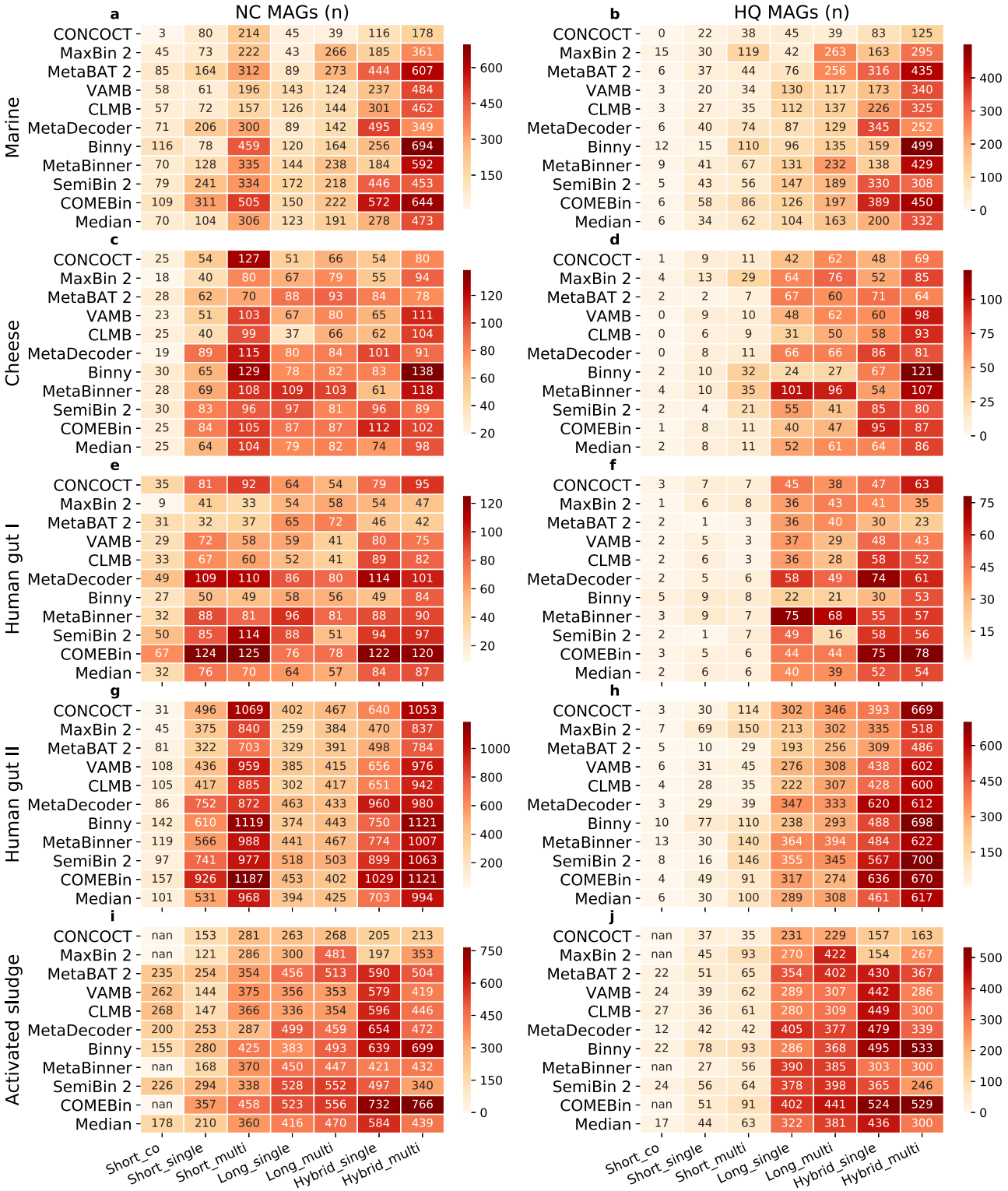

Fig. 1 | Number of NC and HQ MAGs recovered from five real datasets. **a**, **c**, **e**, **g**, **i** The number of NC MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets respectively. **b**, **d**, **f**, **h**, **j** The number of HQ MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets respectively. The description of seven data-binning combinations can be seen in Table 2. "nan" denotes that the corresponding binner failed to complete execution within two weeks based on the computational resources we used.

around 6.3 million contigs (see Section Computational resources for computational resources). Consequently, these binners are ranked last on the activated sludge short-read co-assembly data (see Supplementary Fig. 3e). Moreover, the runtimes and memory usage of each binner across seven data-binning combinations in the activated sludge dataset were assessed (Fig. 2b, c and Supplementary Table 6).

We listed the top three high-performance binners (Table 2) for each data-binning combination based on overall ranking scores. The full rankings of high-performance binners are shown in Fig. 2.

https://doi.org/10.1038/s41467-025-57957-6

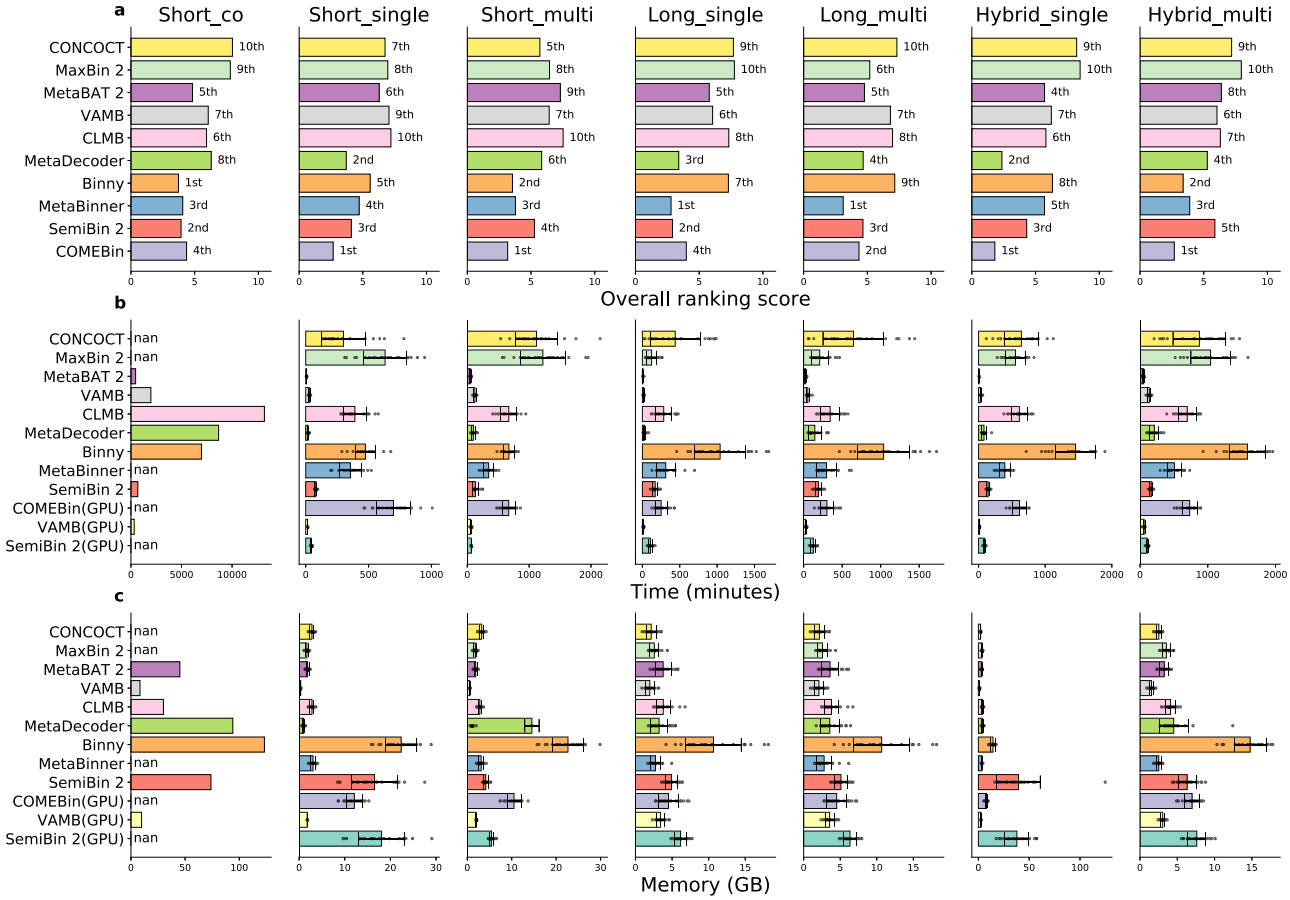

Fig. 2 | Overall ranking scores (see Section Evaluation metrics and ranking score), runtime, and memory usage for each binner across seven data-binning combinations. a Overall ranking scores of each binner across the seven data-binning combinations. The rankings of the binners, based on their overall ranking scores, were annotated. "nan" denotes that CONCOCT, MaxBin 2, MetaBinner, and COMEBin (GPU) cannot complete execution within two weeks on the activated sludge short-read co-assembly data (see Section Computational resources for computational resources). Consequently, these binners are ranked last on the activated sludge short-read co-assembly data (see Supplementary Fig. 3e). **b**, **c** Runtime and memory usage for each binner in the activated sludge dataset. According to the recommendations in the referenced study[20], COMEBin was used in GPU environments. In the short_co data-binning combination, there is a single assembly, with each bar representing the runtime or memory usage for each binner. In the remaining six data-binning combinations, there are 23 assemblies ($N = 23$) per combination, with each bar showing the average runtime or memory usage for each binner. The error bars indicate the standard deviation.

COMEBin, MetaBinner, and SemiBin 2 all ranked within the top five across the seven data-binning combinations. Specifically, COMEBin achieved the highest overall ranking scores for hybrid_multi, hybrid_single, short_multi, and short_single data-binning combinations, while MetaBinner demonstrated the highest overall ranking scores for the long_multi and long_single data-binning combinations. SemiBin 2 performed well in the hybrid_single, long_multi, long_single, and short_single data-binning combinations. Moreover, MetaDecoder excelled in the hybrid_single, long_single, and short_single data-binning combinations, highlighting its suitability for single-sample binning modes. Conversely, Binny achieved superior results in the hybrid_multi, short_multi, and short_co data-binning combinations when the dataset contained a larger number of samples (e.g., 15 or more) but performed poorly in long-read data and the single-sample binning mode. Additionally, the microbial diversity within the bins recovered by the binners with the highest overall ranking scores was further investigated. The results suggested that these binners tend to recover extensive microbial diversity (see Supplementary Note 1; Supplementary Table 7 and Supplementary Figs. 4 and 5).

The time and memory requirements for different binners across seven data-binning combinations exhibit substantial differences (Fig. 2b, c and Supplementary Table 6). According to the recommendations in the referenced study[20], COMEBin was used in GPU environments. Additionally, CONCOCT, MaxBin 2, MetaBinner, and COMEBin (GPU) did not complete execution within two weeks on the activated sludge dataset short-read co-assembly data, which included around 6.3 million contigs (see Section Computational resources for computational resources). MetaBAT 2 was the fastest, completing in 463 minutes, followed by SemiBin2, which took 681 minutes. Moreover, VAMB, CLMB, and MetaBAT 2 exhibited efficient memory usage, consuming 8.34, 30, and 45.18 GB, respectively. In the other six data-binning combinations, MetaBAT 2, VAMB, and MetaDecoder demonstrated outstanding efficiency in both time and memory utilization. Therefore, we listed MetaBAT 2, VAMB, and MetaDecoder as efficient binners (Table 2). The MQ, NC, and HQ MAGs per minute, along with the MQ, NC, and HQ MAGs per GB of memory for each binner on the activated sludge dataset, are presented in Supplementary Tables 8–10 to assess both performance and efficiency.

Identifying the optimal bin-refinement tool

The bin-refinement tools, DAS Tool, MetaWRAP, and MAGScoT were utilized to refine the MAGs recovered by the top three high-performance binners for each data-binning combination listed in Table 2. The numbers of MQ, NC, and HQ MAGs recovered by DAS Tool, MetaWRAP, and MAGScoT across seven data-binning combinations are presented in Fig. 3. Each row of Fig. 3 presents three subplots,

https://doi.org/10.1038/s41467-025-57957-6

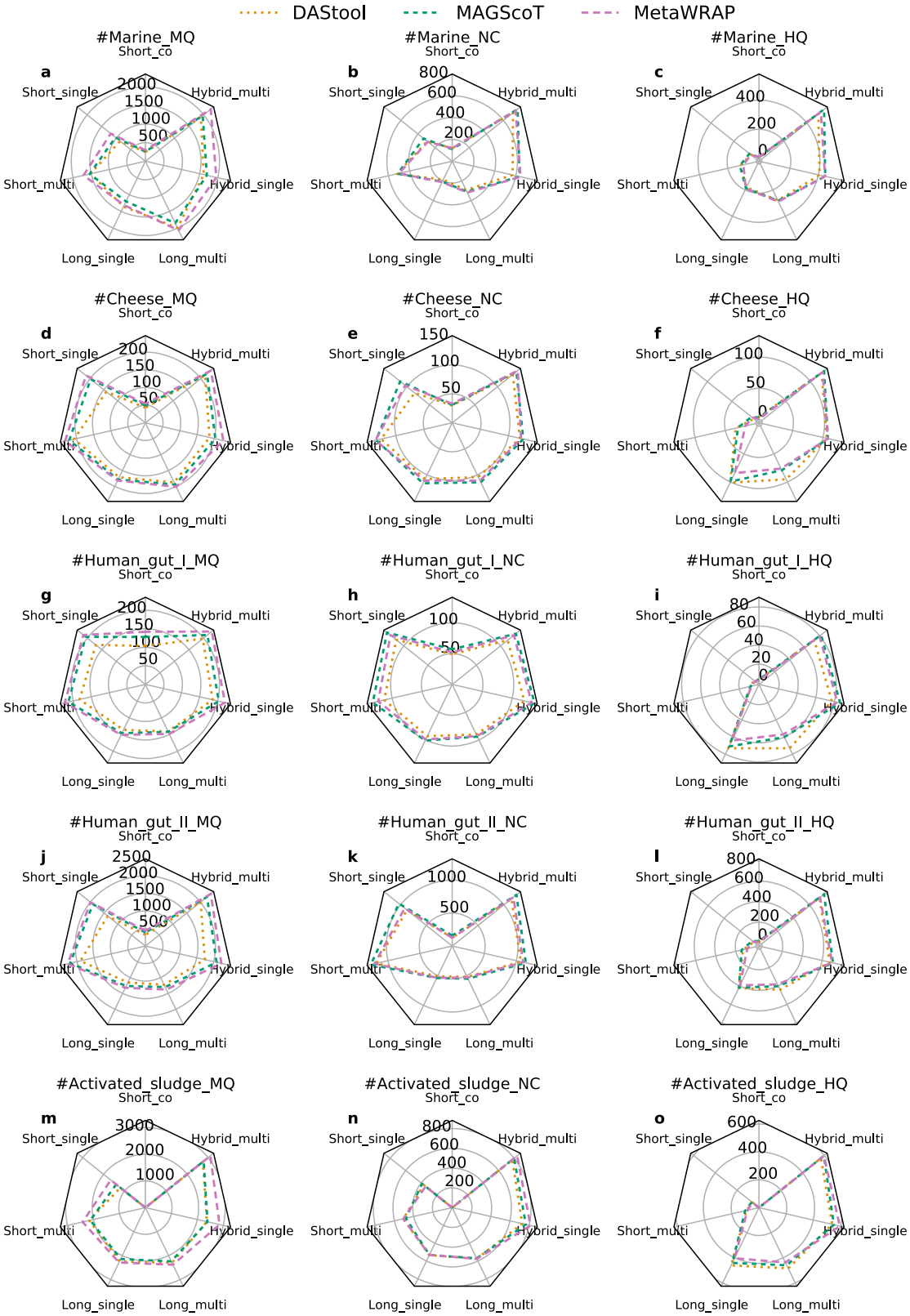

Fig. 3 | Comparison of bin-refinement tool performance across seven data-binning combinations. a, **d**, **g**, **j**, **m** The number of MQ MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets, respectively. **b**, **e**, **h**, **k**, **n** The number of NC MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets, respectively. **c**, **f**, **i**, **l**, **o** The number of HQ MAGs recovered from marine, cheese, human gut I, human gut II and activated sludge datasets, respectively.

illustrating the number of bins across three quality categories (MQ, NC, and HQ) recovered by bin-refinement tools in the seven data-binning combinations for a given environment.

MetaWRAP consistently outperformed both DAS Tool and MAGScoT in the number of MQ MAGs recovered across different datasets (Fig. 3a, d, g, j, and m). Additionally, MAGScoT exhibited the best overall performance in recovering NC MAGs (Fig. 3b, e, h, k, and n). However, none of the three methods showed a remarkable advantage in recovering HQ MAGs. (Fig. 3c, f, i, l, and o).

To select the optimal bin-refinement tool, we calculated the overall ranking scores (see Section Evaluation metrics and ranking score) of all tools based on the number of MQ, NC, and HQ MAGs they retrieved for each data-binning combination in five datasets. Subsequently, we averaged these scores across the seven data-binning combinations for each algorithm. MetaWRAP achieved the best average rank score, while MAGScoT achieved a comparable rank score (Supplementary Table 11). However, MetaWRAP utilized approximately 30 GB of memory across all data-binning combinations, whereas both DAS Tool and MAGScoT required less than 3 GB of memory under the same conditions. (Supplementary Table 6). Additionally, MetaWRAP took more than 10 times longer compared to DAS Tool and MAGScoT (Supplementary Table 6). Therefore, the results obtained from MAGScoT were employed for further analysis. Additionally, we compared MetaBAT 2 with the refined results of MAGScoT in recovering NC and HQ MAGs for each data-binning combination across five datasets (Supplementary Fig. 6), MAGScoT exhibited an average improvement of 59%, 54%, 166%, 99%, 34% more NC MAG on marine, cheese, human gut I, human gut II and activated sludge datasets, respectively, compared to MetaBAT 2. Similar trends were observed in the recovery of HQ MAGs. To illustrate the contributions of the top three high-performance binners, all NC bins recovered by these binners under each data-binning combination in the marine dataset were dereplicated at the species level (see Supplementary Fig. 7). The results demonstrated that the top-ranked binner recovered the most unique species across six data-binning combinations. Moreover, these NC species were annotated using the Genome Taxonomy Database toolkit (GTDB-Tk, version 2.4.0) with the GTDB release r220 (see Supplementary Fig. 8). The results showed that the top-ranked binner revealed highest number of known genus, family, order, and class across six data-binning combinations.

Multi-sample binning recovers extensive species and strains across diverse data types

To compare different data-binning combinations on the marine dataset, the refined NC and HQ MAGs obtained from each data-binning combination underwent separate dereplication procedures (see Section Dereplication and phylogenetic analysis of MAGs) at both the species and strain levels (Fig. 4 and Supplementary Tables 12 and 13).

Compared to short-read single-sample binning, short-read multi-sample binning demonstrated a 41% increase in the number of species (135 versus 96) for NC MAGs and a 43% increase (30 versus 21) for HQ MAGs (Fig. 4). Additionally, there was a 51% increase in the number of strains (229 versus 152) for NC MAGs and a 71% increase (48 versus 28) for HQ MAGs. Moreover, multi-sample binning exhibited similar improvement in long-read and hybrid data. It is worth noting that, in both multi-sample and single-sample binning, HQ MAGs recovered from hybrid data contained the highest number of species and strains, followed by long-read data, and lastly, short-read data. In short-read data, the HQ MAGs from co-assembly binning exhibited the lowest diversity in terms of both species and strains compared to single-sample and multi-sample binning.

Dereplication was also performed on the MAGs recovered from different binning modes within specific types of sequencing data to identify the overlap of species and strains (Supplementary Fig. 9). For short-read, long-read, and hybrid data, multi-sample binning

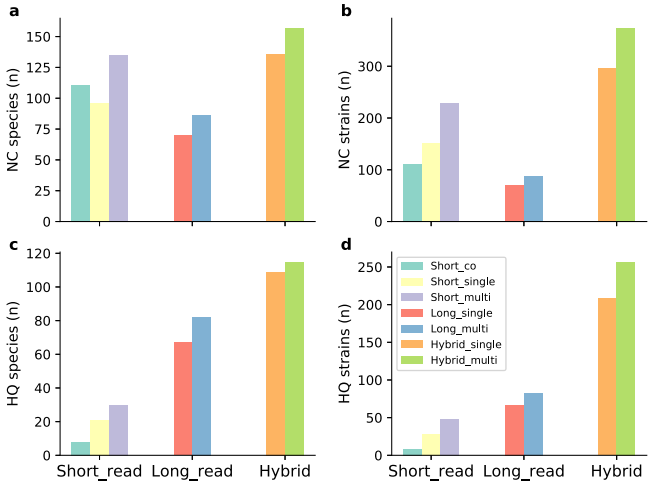

Fig. 4 | Number of species and strains recovered from the marine dataset. **a**, **b** Number of species or strains after dereplication of refined NC MAGs across seven data-binning combinations. **c**, **d** Number of species or strains after dereplication of refined HQ MAGs across seven data-binning combinations.

demonstrated a greater diversity of unique species and strains in recovered NC and HQ MAGs compared to single-sample binning. Additionally, short-read co-assembly binning exhibited the fewest unique species and strains in recovered HQ MAGs compared to short-read single-sample and multi-sample binning. These findings were consistent with the results presented in Fig. 4.

Furthermore, the marine NC and HQ MAGs recovered by the top three high-performance binners were dereplicated at both the species and strain levels for comparison (Supplementary Fig. 10a–d). The results demonstrated that the top-ranked binner recovered the highest number of NC species and strains across the six data-binning combinations. Specifically, the top-ranked binner recovered an average of 22.32% and 39.87% more species across all data-binning combinations, compared to the second- and third-ranked binners, respectively. Similar trends were observed in the recovery of HQ species and strains.

Multi-sample binning reveals extensive potential ARG hosts and BGCs across diverse data types

Antibiotic resistance genes (ARGs) are increasingly acknowledged as global threats to human health[36–39]. In this study, potential hosts of ARGs were identified from the marine refined NC MAGs at the strain-level (NC strains) obtained through seven data-binning combinations (Supplementary Fig. 11a). To enhance the reliability of identifying potential hosts for ARGs, stringent criteria were applied. Potential ARG hosts were defined as NC strains that contained at least one ARG, and the length of the contig associated with the ARG was required to be longer than 10kb[39]. Multi-sample binning exhibited remarkable superiority over single-sample binning by identifying 30%, 22%, and 25% more potential ARG hosts on short-read, long-read, and hybrid data, respectively (Supplementary Fig. 11a). Moreover, ARG hosts identified by multi-sample binning exhibited a greater number of multi-resistance features (see Section Annotation of ARGs and BGCs) compared to those identified by other binning modes (Supplementary Fig. 11b). This highlights the effectiveness of multi-sample binning in identifying potential ARG hosts across diverse data types, thereby facilitating more accurate microbial risk assessments.

Moreover, microbial secondary metabolites have been demonstrated to play a highly significant role in the process of drug discovery and development[40]. BGCs contain the genes responsible for the biosynthesis of these secondary metabolites[41,42]. Here we predicted potential BGCs from the marine NC strains obtained through seven

data-binning combinations (Supplementary Fig. 11c). Similar to the results presented in Supplementary Fig. 11a, multi-sample binning demonstrated superior performance over single-sample binning by identifying 54%, 24%, and 26% more potential BGCs on short-read, long-read, and hybrid data, respectively (Supplementary Fig. 11c). To assess the novelty, the predicted BGCs were subsequently compared to the BiG-FAM[43] database, which includes 1,225,071 known BGCs, using BiG-SLiCE[44] with a threshold of 900[44,45]. The results showed that multi-sample binning identified 36%, 28%, and 22% more potential novel BGCs than single-sample binning in short-read, long-read, and hybrid data, respectively (Supplementary Fig. 11d). This finding illustrated the advantage of multi-sample binning in potential secondary metabolite discovery across diverse data types.

The potential BGCs and ARG hosts were also identified from marine NC strains recovered by the top three high-performance binners (Supplementary Fig. 10e–h). The NC strains recovered by the top-ranked binner exhibited an average of 23.23% and 38.76% more potential BGCs across all data-binning combinations, compared to the second- and third-ranked binners, respectively (Supplementary Fig. 10e). Moreover, the NC strains recovered by the top-ranked binner demonstrated the highest number of potential novel BGCs across six data-binning combinations (Supplementary Fig. 10f). For instance, in the hybrid_single data-binning combination, the top-ranked, second-ranked, and third-ranked binners correspond to COMEBin, MetaDecoder, and SemiBin2, respectively. In this case, the NC strains recovered by COMEBin demonstrated 13.1% and 29.78% more potential novel BGCs than MetaDecoder and SemiBin2, respectively. Similar results were observed in identifying the ARG host from the NC strains recovered by the top three high-performance binners (Supplementary Fig. 10g, h).

Discussion

Metagenomic binning enables the recovery of MAGs without the need for cultivation. MAGs offer vast potential for unraveling the complexities of microbial diversity and ecology. Over the past decade, several benchmarking studies have been proposed. For instance, CAMI (2017)[27] and CAMI II (2022)[28] provided evaluation metrics and simulated benchmark datasets, engaging the global developer community to benchmark their methods. Particularly, they evaluated the performance of metagenomic binning methods on short-read simulated datasets. Additionally, Yue et al.[30] benchmarked metagenomic binning tools on CAMI and real short-read datasets. Recently, Jia et al.[29] benchmarked eight binners on simulated and real gut metagenomics datasets with short-read, long-read and metaHiC sequencing datasets. Moreover, Mattock et al.[26] compared single-sample and multi-sample binning modes on real short-read datasets using MetaBAT 2. In comparison to previous benchmark studies, our work differs in the following aspects: (1) It covers all three binning modes: co-assembly, single-sample and multi-sample binning. (2) It covers short-read, long-read, and hybrid data collected from real samples in various environments. (3) It employs the latest genome quality assessment tool, CheckM 2. (4) It includes both stand-alone binners and bin-refinement tools. (5) It includes recently published binners, such as COMEBin, SemiBin 2 and MetaDecoder.

In this study, we introduced seven data-binning combinations by integrating short-read, long-read, and hybrid data with co-assembly, single-sample, and multi-sample binning, respectively. Subsequently, a comprehensive benchmarking study was conducted to thoroughly evaluate 13 metagenomic binning tools across these seven data-binning combinations based on five real-world datasets with mNGS, PacBio HiFi, and Oxford Nanopore sequencing data. The benchmark results consistently demonstrated that multi-sample binning outperformed single-sample binning across short-read, long-read and hybrid data. Moreover, we observed that short-read data had limited capability in recovering HQ MAGs compared to long-read and hybrid data. Additionally, hybrid data, which combined the strengths of short-read and long-read data, demonstrated the best overall performance.

We then identified the top three high-performance binners for each data-binning combination and listed the overall efficient binners (see Table 2). Three bin-refinement tools were compared by integrating the outputs of the top three binners for each data-binning combination. The results obtained from MAGScoT were employed for further analyses. Dereplication was performed at both the species and strain levels to compare the effectiveness of different binning modes across diverse data types. The results revealed that the multi-sample binning of short-read, long-read, and hybrid data demonstrated a marked enhancement in recovering species and strain diversity compared to the single-sample binning. ARGs and BGCs were also identified in NC strains obtained from seven data-binning combinations. The results demonstrated that employing multi-sample binning led to the detection of the highest number of potential ARG hosts and BGCs in short-read, long-read, and hybrid data. This indicated that utilizing multi-sample binning can facilitate microbial risk assessments and potential secondary metabolite discovery. Therefore, when there are sufficient computational resources and budget, hybrid data and multi-sample binning are optimal choices for users.

Several components could contribute to binning performance. Firstly, deep learning as a powerful tool could effectively integrate heterogeneous features such as tetranucleotide frequency and contig coverage. For instance, the recently developed COMEBin[20] and SemiBin 2[19] utilize contrastive learning to generate HQ feature embeddings that facilitate contig clustering. Secondly, ensemble strategies can enhance the robustness of binning methods. SemiBin 2 employs an ensemble-based DBSCAN algorithm for long-read data, COMEBin utilizes different parameters during clustering and outputs the best result, and MetaBinner[15] applies a two-stage ensemble strategy to produce the final result. However, employing an ensemble strategy requires additional computational resources. Finally, single-copy genes could provide valuable information to guide binning. For instance, MaxBin2[9], MetaDecoder[14], SemiBin 2, and COMEBin leverage single-copy genes for clustering initialization, generating prior probabilities or estimating component binning results to enhance binning performance.

None of the binners consistently performed well across all seven data-binning combinations. It can be attributed to the fact that most binners are specifically designed for short-read data, whereas the assembled contigs from long-read and hybrid data tend to be longer compared to those from short-read data. Moreover, the number of assembled contigs from long-read data is fewer than that from short-read and hybrid data. Additionally, when performing multi-sample and co-assembly binning, the dimensionality of coverage features increases with the growing number of samples. Integrating the strengths of existing binners could lead to the development of a more robust binner suitable for various scenarios. For instance, the representations obtained by deep learning methods such as COMEBin, VAMB, and SemiBin2 can be used as input for clustering methods employed by high-performance binning methods like MetaBinner and MetaDecoder. Specifically, obtaining HQ representations and clustering can be regarded as two distinct modules. These modules from different methods can be integrated to build an extended framework. Different representation or clustering methods can be chosen depending on the data type. For instance, highly efficient modules may be prioritized for large datasets, while HQ modules may be preferred for small datasets.

Although these state-of-the-art metagenomic binning tools were comprehensively compared under seven data-binning combinations, there are limitations in our study. Specifically, our benchmark utilized short-read and long-read sequencing datasets. However, we did not explore additional sequencing techniques, like metagenomic Hi-C, as it requires exclusive Hi-C-based binners. (e.g., bin3C[46], HiCBin[47], and MetaCC[48]). Furthermore, given that CheckM 2[31] is currently the

primary method for assessing the quality of MAGs[6], we utilized it to evaluate the completeness and contamination of MAGs. Although CheckM 2 reliably assesses the quality of archaeal and bacterial genomes, it does not evaluate the quality of genomes from other lineages, such as eukaryotes.

Methods

Benchmarking datasets and preprocessing

Five real metagenomic datasets[29,33,34,49,50] were used in this study based on the following considerations: (1) The datasets were sequenced utilizing both short-read and long-read technologies, thereby meeting the requirements for our benchmarking across seven data-binning combinations. (2) The datasets were collected from diverse environments, enhancing data diversity and ensuring more reliable benchmark results. (3) The datasets were publicly available, ensuring the feasibility of benchmarking and reproducibility of the study. Specifically, the marine dataset consisted of 30 mNGS samples and 30 PacBio HiFi samples[34]. The cheese dataset comprised 15 mNGS samples and 15 PacBio HiFi samples[33]. The human gut I dataset included three mNGS samples and three PacBio HiFi samples[29]. The human gut II dataset contained 30 mNGS samples and 30 Oxford Nanopore samples[50]. The activated sludge dataset comprised 23 mNGS samples and 23 Oxford Nanopore samples[49].

FastQC[51] (version 0.12.1) and MultiQC[52] (version 1.16) were employed to check the quality of Illumina and PacBio HiFi reads, while Nanopore reads were quality-controlled using Nanoplot[53] (version 1.42.0). mNGS samples were preprocessed with Fastp (version 0.23.4)[54] using the following parameters -q 20, - -length_required 100, - -low_complexity_filter. Adapter and barcode sequences in Nanopore samples were trimmed using qcat (version 1.1.0) with the parameters - -trim, - -detect-middle. The trimmed Nanopore samples were then processed with Filtlong (version 0.2.1) to eliminate low-quality and short sequences, using the options - -min_length 4000 and - -min_mean_q 80. Subsequently, residual adapters and barcodes in the Nanopore reads were detected using Porechop (version 0.2.4) with the - -min_split_read_size 4000 option, followed by a final check of Filtlong (version 0.2.1) with the aforementioned options. Moreover, human reads from the human gut samples were removed by mapping sequencing reads to the hg38 reference genome using Bowtie2 (version 2.5.1)[55].

Metagenome assembly and alignment

The short-read co-assembly and single-sample assembly were performed using MEGAHIT (version 1.2.9)[56]. Flye (version 2.9.2)[57] and OPERA-MS (version 0.9.0)[35] were utilized for long-read single-sample and hybrid single-sample assembly, respectively (Supplementary Tables 1–4). Specifically, hybrid single-sample assembly means that each short-read sample and its corresponding long-read sample were assembled together. The resulting hybrid assembly was then polished using Pilon[58] if the long-read sample originates from the Nanopore sequencing platform. Contigs longer than 1kb were selected for subsequent binning. Bowtie2 (version 2.5.1)[55] and minimap2 (version 2.24-r1171, -x map-hifi)[59] were employed to align the short reads and long reads to the contigs, respectively. In hybrid data binning, if the hybrid assembly was generated from mNGS and HiFi samples, both short-read and long-read alignments were used as input for the binning processes. SAMtools (version 1.3.1)[60] was utilized to sort the alignment files.

Contig binning

Metagenomic binning contains three modes: co-assembly, single-sample, and multi-sample binning. In co-assembly binning, all sequencing samples are initially assembled together. The assembled contigs are then binned with coverage profiles derived from the reads across all sequencing samples. On the other hand, both single-sample and multi-sample binning involve independent assembly for each sequencing sample. The key difference lies in the method of calculating the coverage profiles: single-sample binning calculates coverage from a single sample, whereas multi-sample binning requires an all-against-all comparison to calculate coverage from all samples. The assembled sample-specific contigs are then binned for each sample independently. In this study, all binning methods utilized the same multi-sample binning mode as detailed in this section.

Contig binning was performed using CONCOCT (version 1.1.0)[8], MaxBin 2 (version 2.2.7)[9], MetaBAT 2 (version 2.15)[10], VAMB (version 3.0.9)[16], CLMB (version 1.0.0)[17], MetaDecoder (version 1.0.16)[14], Binny (version 2.2.15)[12], MetaBinner (version 1.4.4)[15], SemiBin 2 (version 1.5.1)[19], and COMEBin[20] (version 1.0.3). DAS Tool (version 1.1.6)[23], MetaWRAP (version 1.3.2)[22], and MAGScoT (version 1.0.0)[24] were used to refine the outputs of multiple binning tools (Table 1).

Evaluation metrics and ranking score

According to guidelines of CAMI II[28] and MIMAG[32], three criteria were used for assessing the quality of MAGs, including HQ (completeness > 90%, contamination < 5% and presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs), NC (completeness > 90% and contamination < 5%), "moderate or higher" quality (MQ, completeness > 50% and contamination < 10%). To evaluate the completeness and contamination levels of MAGs, CheckM 2[31] (version 1.0.2) was utilized. The presence of transfer RNA (tRNA) genes was detected using Aragorn[61] (version 1.2.41), while Barrnap (version 0.9, https://github.com/tseemann/barrnap) was employed to predict the location of 5S, 16S, and 23S rRNA genes.

The overall ranking score for each binner was calculated by integrating their performance across all datasets based on MQ, NC, and HQ metrics. Firstly, we calculated the rank_MQ, rank_NC, and rank_HQ for all binners based on each assembly, where the top-ranked binner for a given metric in a sample received a rank value of 1, the second-ranked received a rank value of 2, and so on. Secondly, we averaged the rank_MQ, rank_HQ, and rank_NC of all assemblies within each dataset to obtain the ranking score for each dataset, as follows:

$$\text{Ranking score}_{\text{dataset}} = \frac{1}{3}\left(\frac{1}{n}\sum_{i=1}^{n}\text{rank\_MQ}_i + \frac{1}{n}\sum_{i=1}^{n}\text{rank\_NC}_i + \frac{1}{n}\sum_{i=1}^{n}\text{rank\_HQ}_i\right) \quad (1)$$

where n denotes the number of samples in the dataset. Finally, we calculated the overall ranking score by averaging the ranking scores across all datasets, as follows:

$$\text{Overall ranking score} = \frac{1}{N}\sum_{i=1}^{N}\text{ranking score}_{\text{dataset}_i} \quad (2)$$

where N denotes the number of real datasets.

Dereplication and phylogenetic analysis of MAGs

To dereplicate the MAGs at the species (95% nucleotide identity (ANI) threshold) and strain levels (99% ANI threshold), the software dRep[62] (version 3.4.3) was used with the options -p 32, -nc 0.6 -sa 0.95 or -sa 0.99. MAGs were annotated using the Genome Taxonomy Database toolkit[63] (GTDB-Tk, version 2.4.0) with the GTDB release r220. The phylogenetic trees were built using IQ-TREE (version 2.3.5) with the option -m LG+R4 and visualized with iTOL v7[64].

Annotation of ARGs and BGCs

ARGs within MAGs were predicted using RGI (version 6.0.2) with default parameters, relying on the Comprehensive Antibiotic Research Database (CARD version 3.2.7)[65]. A potential host for ARGs was identified as non-redundant NC MAGs at the strain-level (NC strains) that contained at least one ARG. Additionally, the length of the contig

associated with the ARG was required to be over 10kb, ensuring a rigorous selection process[39]. A potential ARG host with multi-resistance features indicates that the host carries multiple types of ARGs[66]. Potential BGCs in NC strains were annotated utilizing antiSMASH[67] (version 6.1.1). The novelty of BGCs was assessed using BiG-SLiCE[44] (version 1.1.1) in query mode, with a comparison against 1,225,071 BGCs from the BiG-FAM[43] database and a threshold of 900[44,45]. BGCs with a BiG-SLiCE distance greater than 900 were identified as potential novel BGCs.

Computational resources

We evaluated the runtime and memory usage across seven data-binning combinations on the activated sludge dataset (see Supplementary Table 6). In GPU mode, VAMB (GPU), Semibin 2 (GPU), and COMEBin (GPU) were run with 16 threads across seven data-binning combinations on a machine with two Intel(R) Xeon(R) Silver 4110 CPUs @ 2.10GHz and a GeForce GTX 1080Ti GPU. In CPU-only mode, all tools were run with 32 threads for the short-read co-assembly data-binning combination and 16 threads for the other six data-binning combinations, on a machine with two Hygon C86 7285 32-core Processors.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the datasets used in this study are publicly available. The marine dataset, comprising 30 Illumina samples and 30 PacBio HiFi samples, is available in the NCBI BioProject under the accession code PRJEB52999. The cheese dataset, consisting of 15 Illumina samples and 15 PacBio HiFi samples, is available on the NCBI BioProject under the accession code PRJNA778418. The human gut I dataset, including three Illumina samples and three PacBio HiFi samples, is available in the National Genomics Data Center (NGDC: https://ngdc.cncb.ac.cn/) under accession code PRJCA007414. The human gut II dataset, containing 30 illumina samples and 30 Oxford Nanopore samples, is available in the NCBI BioProject under the accession code PRJNA820119. The activated sludge dataset, comprising 23 Illumina samples and 23 Oxford Nanopore samples, is available on the NCBI BioProject under the accession code PRJNA629478. Run accessions of all the samples are given in Supplementary Table 5. The MQ, NC, and HQ bins obtained with the refinement tool, across seven data-binning combinations and five real datasets, along with the corresponding CheckM2 output files, have been uploaded to https://zenodo.org/records/14535874.

Code availability

We provided a user-friendly metagenomic binning wrapper suite that comprises two efficient binners (MetaBAT 2, MetaDecoder), two high-performance binners (MetaBinner, COMEBin), and a fast bin-refinement tool MAGScoT. Commands and scripts of this study are available at https://github.com/htaohan/databinning. The scripts are also archived on Zenodo at https://doi.org/10.5281/zenodo.14906397.

References

1. Jansson, J. K. Microorganisms, climate change, and the sustainable development goals: progress and challenges. *Nat. Rev. Microbiol.* **21**, 622–623 (2023).
2. Prosser, J. I. et al. The role of ecological theory in microbial ecology. *Nat. Rev. Microbiol.* **5**, 384–392 (2007).
3. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6**, 805–814 (2005).
4. Zeng, S. et al. A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat. Commun.* **13**, 5139 (2022).
5. Albertsen, M. Long-read metagenomics paves the way toward a complete microbial tree of life. *Nat. Methods* **20**, 30–31 (2023).
6. Malmstrom, R. R. Quality MAGnified. *Nat. Rev. Microbiol.* **21**, 771 (2023).
7. Nayfach, S. et al. A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
8. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
9. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
10. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, 7359 (2019).
11. Zhu, X. & Ghahramani, Z. Learning from Labeled and Unlabeled Data with Label Propagation. Report No. CMU-CALD-02-107 (Carnegie Mellon University, 2002).
12. Hickl, O., Queirós, P., Wilmes, P., May, P. & Heintz-Buschart, A. binny: an automated binning algorithm to recover high-quality genomes from complex metagenomic datasets. *Brief. Bioinforma.* **23**, 431 (2022).
13. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining* (eds Pei, J. et al.) (Springer, 2013).
14. Liu, C.-C. et al. Metadecoder: a novel method for clustering metagenomic contigs. *Microbiome* **10**, 1–16 (2022).
15. Wang, Z., Huang, P., You, R., Sun, F. & Zhu, S. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol.* **24**, 1 (2023).
16. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).
17. Zhang, P., Jiang, Z., Wang, Y. & Li, Y. CLMB: deep contrastive learning for robust metagenomic binning. In *Proc. 26th Annual International Conference on Research in Computational Molecular Biology: RECOMB 2022*. San Diego, CA, USA, May 22–25, 2022, (ed Pe'er, I.) 326–348 (Springer, 2022).
18. Pan, S., Zhu, C., Zhao, X.-M. & Coelho, L. P. A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nat. Commun.* **13**, 2326 (2022).
19. 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**, 21–29 (2023).
20. Wang, Z. et al. Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat. Commun.* **15**, 585 (2024).
21. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
22. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP–a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 1–13 (2018).
23. Sieber, C. M. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
24. Rühlemann, M. C., Wacker, E. M., Ellinghaus, D. & Franke, A. MAGScoT: a fast, lightweight and accurate bin-refinement tool. *Bioinformatics* **38**, 5430–5433 (2022).
25. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 1–11 (2016).

26. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat. Methods* **20**, 1170–1173 (2023).
27. Sczyrba, A. et al. Critical assessment of metagenome interpretation–a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
28. Meyer, F. et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
29. Jia, L. et al. A survey on computational strategies for genome-resolved gut metagenomics. *Brief. Bioinform.* **24**, bbad162 (2023).
30. Yue, Y. et al. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinforma.* **21**, 1–15 (2020).
31. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
32. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
33. Saak, C. C. et al. Longitudinal, multi-platform metagenomics yields a high-quality genomic catalog and guides an in vitro model for cheese communities. *Msystems* **8**, 00701–22 (2023).
34. Orellana, L. H., Krüger, K., Sidhu, C. & Amann, R. Comparing genomes recovered from time-series metagenomes using long-and short-read sequencing technologies. *Microbiome* **11**, 105 (2023).
35. Bertrand, D. et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* **37**, 937–944 (2019).
36. Allen, H. K. et al. Call of the wild: antibiotic resistance genes in natural environments. *Nat. Rev. Microbiol.* **8**, 251–259 (2010).
37. Manaia, C. M. Assessing the risk of antibiotic resistance transmission from the environment to humans: non-direct proportionality between abundance and risk. *Trends Microbiol.* **25**, 173–181 (2017).
38. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
39. Zhang, Z. et al. Assessment of global health risk of antibiotic resistance genes. *Nat. Commun.* **13**, 1553 (2022).
40. Pan, R., Bai, X., Chen, J., Zhang, H. & Wang, H. Exploring structural diversity of microbe secondary metabolites using OSMAC strategy: A literature review. *Front. Microbiol.* **10**, 294 (2019).
41. Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
42. Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief. Bioinforma.* **20**, 1103–1113 (2019).
43. Kautsar, S. A., Blin, K., Shaw, S., Weber, T. & Medema, M. H. BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.* **49**, 490–497 (2021).
44. Kautsar, S. A., Hooft, J. J., Ridder, D. & Medema, M. H. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience* **10**, 154 (2021).
45. Du, R., Xiong, W., Xu, L., Xu, Y. & Wu, Q. Metagenomics reveals the habitat specificity of biosynthetic potential of secondary metabolites in global food fermentations. *Microbiome* **11**, 115 (2023).
46. DeMaere, M. Z. & Darling, A. E. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biol.* **20**, 1–16 (2019).
47. Du, Y. & Sun, F. HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. *Genome Biol.* **23**, 63 (2022).
48. Du, Y. & Sun, F. MetaCC allows scalable and integrative analyses of both long-read and short-read metagenomic Hi-C data. *Nat. Commun.* **14**, 6231 (2023).
49. Singleton, C. M. et al. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat. Commun.* **12**, 2009 (2021).
50. Chen, L. et al. Short-and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nat. Commun.* **13**, 3175 (2022).
51. Andrews, S. et al. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. (Cambridge, United Kingdom, 2010).
52. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
53. De Coster, W., D'hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
54. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
55. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
56. 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).
57. Kolmogorov, M. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).
58. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS* ONE **9**, 112963 (2014).
59. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
60. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 008 (2021).
61. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
62. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
63. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
64. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, 256–259 (2019).
65. Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, 690–699 (2023).
66. Zhao, R. et al. Deciphering the mobility and bacterial hosts of antibiotic resistance genes under antibiotic selection pressure by metagenomic assembly and binning approaches. *Water Res.* **186**, 116318 (2020).
67. Blin, K. et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, 29–35 (2021).
68. Song, W.-Z. & Thomas, T. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics* **33**, 1873–1875 (2017).

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant Nos. U24A20257 and 62272105), ZJ Lab, Shanghai Center for Brain Science and Brain-Inspired Intelligence Technology, and 111 Project (Grant No. B18015).

Author contributions

S.Z. conceived and supervised the project. S.Z., H.H., and Z.W. designed the study and the benchmark framework. H.H. and Z.W. collected the datasets. H.H. implemented the benchmarking pipeline. S.Z., H.H., and Z.W. analyzed the results. H.H. drafted the paper. S.Z. and Z.W. modified the paper. All authors agree to the content of the final paper

Competing interests

All authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57957-6.

Correspondence and requests for materials should be addressed to Shanfeng Zhu.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.