REMAG: recovery of eukaryotic genomes from metagenomic data using contrastive learning

Daniel Gómez-Pérez [1,✉], **Sébastien Raguideau** [1], **Sally Warring** [1], **Robert James** [1], **Falk Hildebrand** [1,2], **and Christopher Quince** [1,2,3,✉]

[1]Earlham Institute, Norwich Research Park, NR4 7UZ Norwich, UK; [2]Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ Norwich, UK; [3]School of Biological Sciences, University of East Anglia, NR4 7TJ Norwich, UK

Metagenome-assembled genomes (MAGs) are central to exploring microbial communities. Yet, despite the relevance of protists and fungi to diverse ecosystems, eukaryotic MAG recovery lags behind that of prokaryotes. A major bottleneck is that most state-of-the-art binning pipelines exclusively rely on prokaryotic single-copy core gene reference databases and are optimized for smaller genomes.
To address this gap, we present REMAG (Recovery of Eukaryotic MAGs), a tool designed to recover high-quality eukaryotic genomes suited for long-read metagenomic data. REMAG leverages fine-tuned HyenaDNA genomic foundation models to efficiently filter eukaryotic contigs. It then employs a dual-encoder Siamese network trained with Barlow Twins contrastive loss to learn a shared embedding space by integrating contig composition and differential coverage. Finally, high-quality bins are extracted using greedy iterative Leiden clustering optimized with eukaryotic single-copy core gene constraints. In benchmarks based on simulated mixed prokaryotic/eukaryotic communities and real datasets of varying sizes and origin, we demonstrate REMAG's ability to recover more near-complete eukaryotic genomes than existing state-of-the-art tools, which often produce highly fragmented eukaryotic bins.
REMAG provides an automated eukaryotic binning method that scales effectively with the increasing size and sequencing depth of metagenomic datasets.

metagenomics | eukaryotic genomes | contrastive learning | genome binning | foundation models | long-read sequencing

Correspondence: ***daniel.gomez-perez@earlham.ac.uk; christopher.quince@earlham.ac.uk***

Introduction

The popularity of metagenomic approaches for analyzing microbial communities has increased steadily in recent years, due to their potential to resolve individual genomes (Quince et al., 2017). Technological advancements that have contributed to this include long-read sequencing (Jain et al., 2016) and increases in sequencing depth from low-input material (Goodwin et al., 2016). Shotgun metagenomic sequencing data allow the building of comprehensive metagenome-assembled genome (MAG) catalogs (Singleton et al., 2026), which can recover high-quality genomes that in some cases exceed the quality of currently available reference genomes (Cuscó et al., 2025). A key step in producing these catalogs is the binning of contigs generated after metagenomic read assembly. Although state-of-the-art binners can recover near-complete bacterial genomes, this is not the case for eukaryotes, which have longer and more heterogeneous genomes (West et al., 2018; Delmont et al., 2022). Unlike bacterial and archaeal genomes, eukaryotic genomes are typically much larger and are characterized by lower gene density, polyploidy, frequent introns and substantial repetitive content (Lynch and Conery, 2003; Elliott and Gregory, 2015). Together with lower coverage, these distinct features present challenges including variable coverage patterns that can confound binners optimized to work with prokaryotic constraints.

Microbial eukaryotes have a substantial impact on the environment, driving a large part of primary production and comprising key pathogens (Parfrey et al., 2011; Falkowski et al., 2008). Despite their relevance, many are not yet sequenced due to challenges related to culturing under laboratory conditions (del Campo et al., 2014; Massana, 2011), and therefore would benefit from recovery through environmental sample collection and metagenomic analysis.

In metagenomic studies, eukaryotes have historically been neglected due to their low coverage and poor recovery. Of all the general-purpose unsupervised binners for metagenomics (CONCOCT, Alneberg et al., 2014; MaxBin, Wu et al., 2014; MetaBAT2, Kang et al., 2019; among others), none explicitly accommodate eukaryotes. Improvements in sequencing technologies and assemblers, as well as the availability of datasets enriched for eukaryotes by size filtration, such as *Tara* Oceans (Sunagawa et al., 2015), have highlighted the need for specialized eukaryotic binners. Some recently developed pipelines for short-read sequencing, such as Eukfinder (Zhao et al., 2025), are dedicated to recovering eukaryotic MAGs (eMAGs). However, they require large reference databases, which leads to slow query times and reference biases. In addition, they scale poorly with increased diversity, such as that found in newly sequenced microbial communities. Therefore, there is a need for a tool that can recover high-quality eMAGs from mixed prokaryotic/eukaryotic communities

with minimal reference input.

A new generation of automated binners, including SemiBin2 (Pan et al., 2023) and COMEBin (Wang et al., 2024), is based on contrastive learning, alongside tools like VAMB (Nissen et al., 2021) that utilize deep learning to improve binning performance. Contrastive learning in particular (Chen et al., 2020; Khosla et al., 2020) is a technique that has shown remarkable results for the problem of binning. Contrastive learning relies on the creation of representations that capture similarity and dissimilarity relationships among contigs (LeCun et al., 2015; Eraslan et al., 2019). This is achieved by training a neural network to place data points from the same category (positive pairs) closer together in the latent space while separating those from different categories (negative pairs). In the context of binning, positive pairs correspond to versions of the same contig (halves in SemiBin2 and random subsegments in COMEBin). Negative pairs are generated randomly in both programs from different contigs assumed to be unrelated.

However, these approaches are primarily optimized for prokaryotes as they rely heavily on single-copy core genes (SCGs) from archaea and bacteria for accurate clustering and do not explicitly account for the complexity of eukaryotic genomes. Following the paradigm that there is no best binner for every case, we have developed REMAG, which extends the concept of contrastive learning for unsupervised binning, focusing exclusively on the recovery of eukaryotes.

Results

Pipeline overview

REMAG implements a seven-stage integrated pipeline for eukaryotic genome recovery from metagenomic data (Figure 1). The pipeline consists of (1) eukaryotic contig filtering using fine-tuned HyenaDNA (Nguyen et al., 2023) classifiers to reduce computational load on non-target sequences and cross-domain contamination; (2) data augmentation of original contigs to generate positive pairs for training; (3) feature extraction generating compositional (tetranucleotide frequencies) and abundance (coverage) features from original and augmented contig fragments; (4) dual-encoder Siamese neural network training with Barlow Twins contrastive loss to learn embeddings from both feature modalities; (5) k-nearest neighbor (k-NN) graph construction from learned fused embeddings; (6) greedy iterative Leiden clustering to extract high-quality bins; and (7) satellite rescue where fragmented bins are merged into larger core bins based on embedding centroid similarity, provided the merger does not significantly increase single-copy core gene (SCG) duplication. The pipeline outputs clusters larger than 500,000 bp in combined length as bins.

Table 1. Recall and F1 score performance of HyenaDNA models on simulated datasets grouped by biome, sequencing technology (Oxford Nanopore Technologies [ONT], Illumina and PacBio) and tool (Tiara (Karlicki et al., 2022), 4-CAC (Pu and Shamir, 2024) and REMAG). Best results for each dataset and technology are bolded.

Dataset	Technology	Tool	Recall	F1 Score
Gut	Illumina	Tiara	0.894	**0.772**
		4-CAC	0.765	0.705
		REMAG	**0.972**	0.515
	PacBio	Tiara	0.975	**0.667**
		4-CAC	0.925	0.597
		REMAG	**1.000**	0.331
	ONT	Tiara	0.949	**0.679**
		4-CAC	0.949	0.632
		REMAG	**1.000**	0.306
Marine	Illumina	Tiara	0.716	0.817
		4-CAC	0.269	0.405
		REMAG	**0.959**	**0.877**
	PacBio	Tiara	0.791	0.882
		4-CAC	0.846	0.915
		REMAG	**0.999**	**0.994**
	ONT	Tiara	0.783	0.877
		4-CAC	0.831	0.906
		REMAG	**0.998**	**0.993**
Soil	Illumina	Tiara	0.854	0.913
		4-CAC	0.486	0.639
		REMAG	**0.985**	**0.939**
	PacBio	Tiara	0.829	0.906
		4-CAC	0.779	0.875
		REMAG	**0.999**	**0.993**
	ONT	Tiara	0.825	0.903
		4-CAC	0.803	0.889
		REMAG	**0.999**	**0.989**
Plant	Illumina	Tiara	0.884	0.937
		4-CAC	0.923	0.959
		REMAG	**0.995**	**0.993**
	PacBio & ONT	Tiara	0.926	0.958
		4-CAC	0.916	0.954
		REMAG	**0.996**	**0.986**

Benchmark on synthetic datasets

We tested REMAG on mixed prokaryotic/eukaryotic communities from different biomes, including human gut, marine, plant-associated and soil for Illumina short reads and two long-read technologies, Oxford Nanopore Technologies (ONT) and PacBio. Each dataset included a variable number of low-coverage eukaryotic genomes mixed with a larger majority of prokaryotic genomes. Using these four datasets, we compared against CONCOCT (Alneberg et al., 2014), SemiBin2 (Pan et al., 2023) and COMEBin (Wang et al., 2024), using identical assembly inputs and default parameters in single-sample and co-assembly approaches.

Eukaryotic contig filtering. As a first step, REMAG filters the assembly for eukaryotic contigs using HyenaDNA

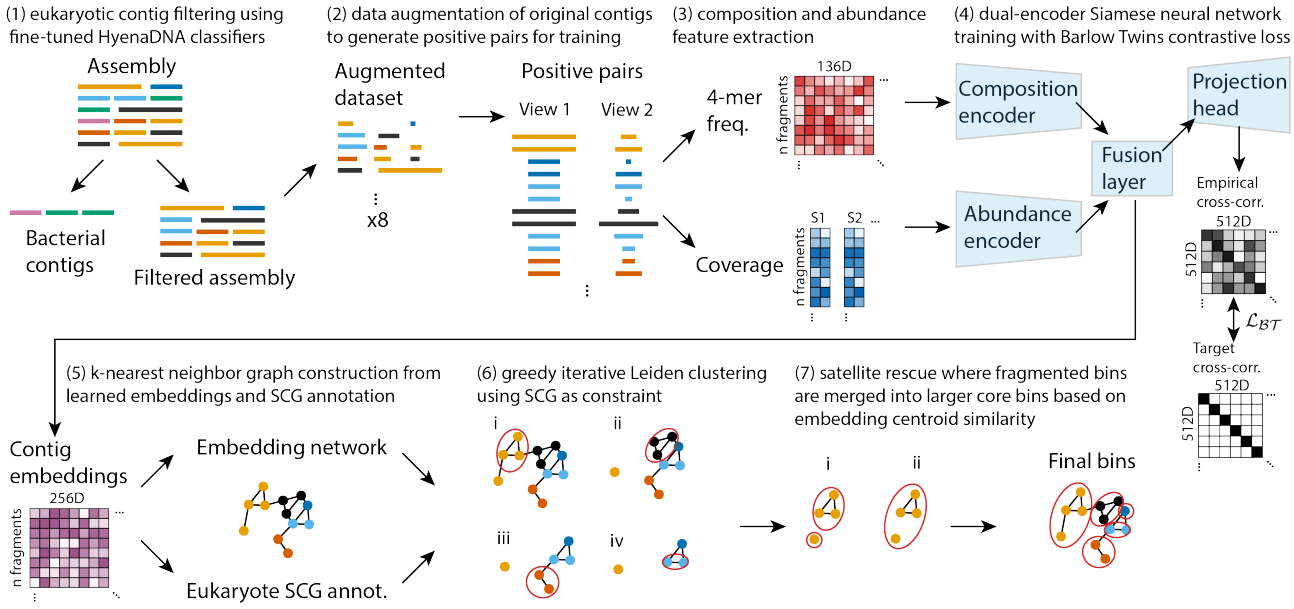

Figure 1. REMAG workflow
Overview of the REMAG pipeline, including (1) eukaryotic contig filtering, (2) and (3) data augmentation and feature extraction, (4) contrastive embedding with Barlow Twins loss ($\mathcal{L}_{\mathcal{BT}}$), and (5) to (7) embedding clustering and bin rescue.

models (Nguyen et al., 2023). This is critical to reduce the size of the input data, improving speed and reducing the potential for spurious bacterial contamination. We benchmarked our filtering approach on the co-assemblies of each of the four simulated datasets by comparing performance to two other dedicated tools for this purpose, Tiara and 4-CAC (Karlicki et al., 2022; Pu and Shamir, 2024). The algorithm for scoring and selecting eukaryotic contigs was optimized for high recall, with the goal of minimal loss of eukaryotic contigs in this step. Indeed, when comparing with other state-of-the-art contig classification tools, REMAG resulted in the highest recall of eukaryotic contigs across all datasets. Despite the higher recall at the cost of lower precision, we observed higher F1 scores across all datasets for REMAG except for the gut dataset, which had a lower F1 compared to other tools (Table 1). This drop in precision is mitigated downstream, as the contrastive learning and Leiden clustering steps effectively isolate false-positive bacterial contigs into separate bins.

Binning on per-sample simulated dataset assemblies. In the per-sample assemblies of the simulated datasets all tools showed low recovery of eMAGs, except for REMAG and CONCOCT on the soil dataset (Supplementary Figure S1). This is likely due to the low coverage of eukaryotes in these datasets, which is expected given that they were designed to reflect real microbial communities where eukaryotes are often underrepresented. These results highlight the importance of co-assembly for recovering eMAGs from metagenomic data, as it allows for increased coverage of low-abundance eukaryotes and improved assembly quality, which are critical for successful binning. Across all datasets, REMAG performed better than other tools: in the PacBio dataset by recovering more high quality (HQ) eMAGs than CONCOCT (9 vs. 5), and in the Illumina and ONT datasets by recovering a larger number of medium quality (MQ) eMAGs (15 vs. 12).

Binning on short-read simulated datasets. In the short-read co-assembly simulated datasets, although being optimized for long-reads, we observed a good performance from REMAG. With seven recovered HQ eMAGs, it tied with CONCOCT, but was able to recover two more MQ eMAGs across all four datasets. While REMAG was outperformed by CONCOCT in the marine and plant-associated datasets, it recovered significantly more HQ and MQ eMAGs in the soil dataset (5 vs. 2; Supplementary Figure S2). Of note, CONCOCT outperforms more modern tools for eukaryotic binning in short-read datasets (7 recovered HQ eMAGs vs. 4 by each SemiBin2 and COMEBin across all datasets). This is likely because while newer binners rely on markers for clustering and refining bins, CONCOCT is agnostic to single-copy markers.

Binning on long-read simulated datasets. In the long-read co-assembly simulated datasets, REMAG performed exceptionally well, consistently recovering more eMAGs across platforms and datasets, both for HQ alone and HQ+MQ. This corresponded to more than double HQ eMAGs across all datasets for both long-read technologies compared to the second-best, COMEBin (20 vs. 9). Additionally, we observed consistently more accurate classification of contigs into eukaryotic bins as measured by a eukaryotic Adjusted Rand Index (eARI) metric (Figure 2). Across both datasets, REMAG had the highest classification accuracy on average (eARI=0.79), while CONCOCT was second with eARI=0.44.
When assessing the performance of REMAG on spe-

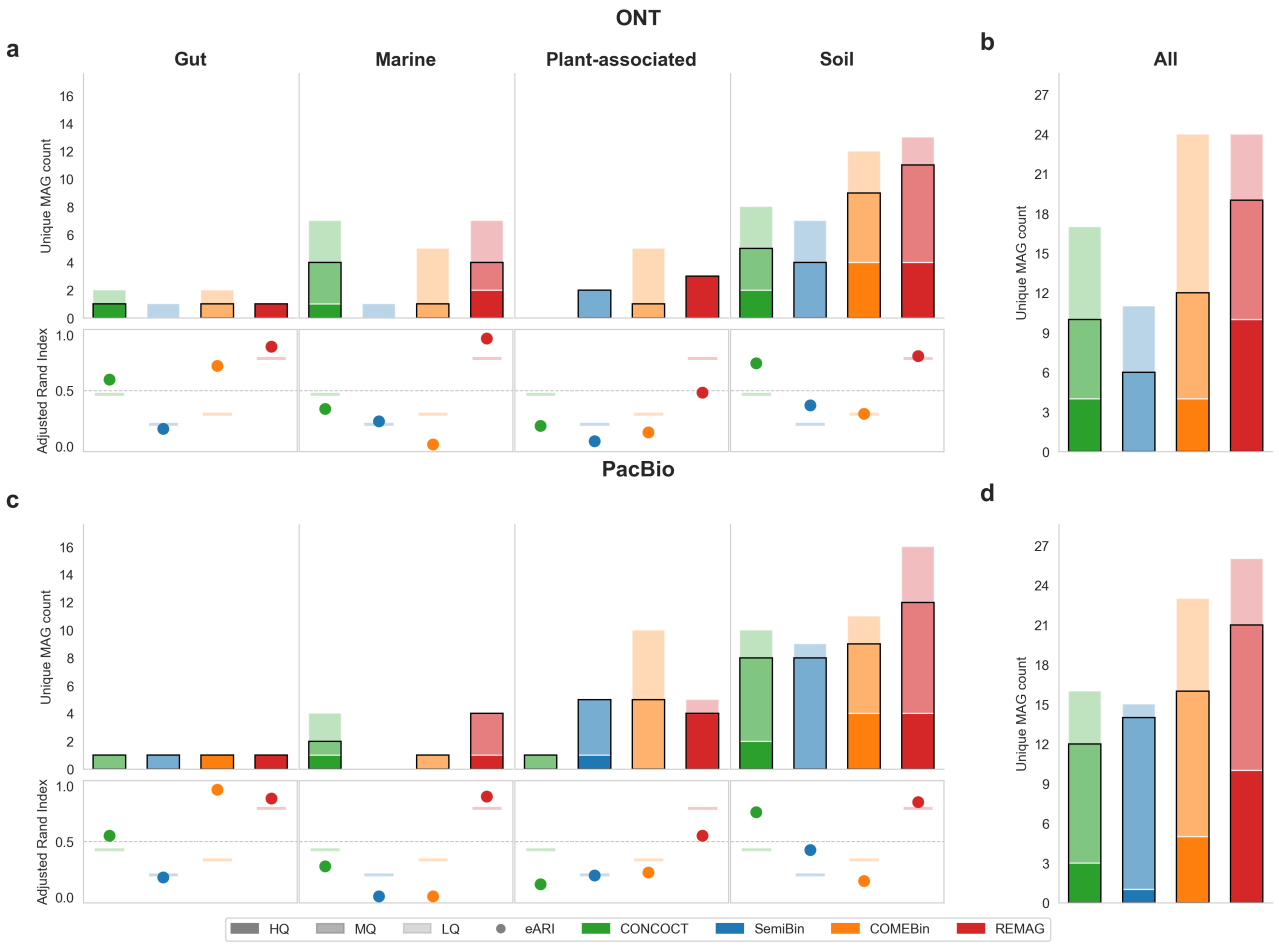

Figure 2. Eukaryotic MAG recovery performance across simulated long-read sequencing datasets
Comparison of binning tools (CONCOCT, SemiBin2, COMEBin, REMAG) on simulated datasets using two long-read sequencing technologies: Oxford Nanopore Technologies (ONT) and PacBio. **a** Unique eukaryotic MAG count and Adjusted Rand Index (ARI) for ONT datasets and **b** sum across all datasets. **c** and **d** show similar plots for PacBio. The eukaryotic ARI (eARI) includes only bins with at least one eukaryotic contig, and the mean across all four datasets is shown as a faint line. HQ: high quality, MQ: medium quality, LQ: low quality.

cific biomes, we noticed a larger improvement compared to other tools on environments with a larger diversity pool, including marine and plant-associated, which both had more than 50 unique eukaryotic genomes in the simulated datasets. Across both datasets, REMAG recovered five times as many HQ eMAGs in the co-assemblies than the second-best tool, CONCOCT (10 vs. 2; Figure 2). In the single-sample assemblies, the recovery difference corresponded to slightly more than double that of CONCOCT (5 vs. 2; Supplementary Figure S1). This demonstrates the scaling potential of REMAG when employed on high-diversity environmental samples.

Computation time. When assessing the runtime compared to the other binners, REMAG was the fastest with an average of 26 min across all the simulated dataset co-assemblies (Supplementary Figure S4), roughly half the runtime of the second fastest tool, CONCOCT (47 min). Although COMEBin showed the second-best performance in the simulated long-read datasets, it was much slower on average, taking 11 h or ≈25 times longer compared to REMAG to complete.

Ablation testing. To evaluate the contribution of different components of REMAG to its overall performance, we performed an ablation study using the simulated datasets. We tested the impact of the following steps by excluding them from otherwise full runs on all simulated datasets: (1) the eukaryotic contig filtering and (2) the satellite bin rescue.

We observed a significant drop in performance when removing the eukaryotic contig filtering step for REMAG. With filtering on, the number of recovered HQ eMAGs across all datasets and sequencing technologies increased by 7 HQ eMAGs (+11.3%), as well as the eARI (+8.6%; Supplementary Figure S5). There were exceptions to the latter, particularly in the short-read datasets. These results highlight the importance of bacterial contig removal for improving the signal-to-noise ratio during eukaryotic binning. Additionally, this step significantly reduced the computational burden of REMAG, resulting in six times faster runtimes on average (26 min vs. 3 h) when filtering is applied (Supplementary Figure S3).

The rescue step also contributed to improved performance, helping to recover three additional MQ eMAGs (+4.8%) across all simulated datasets that would have

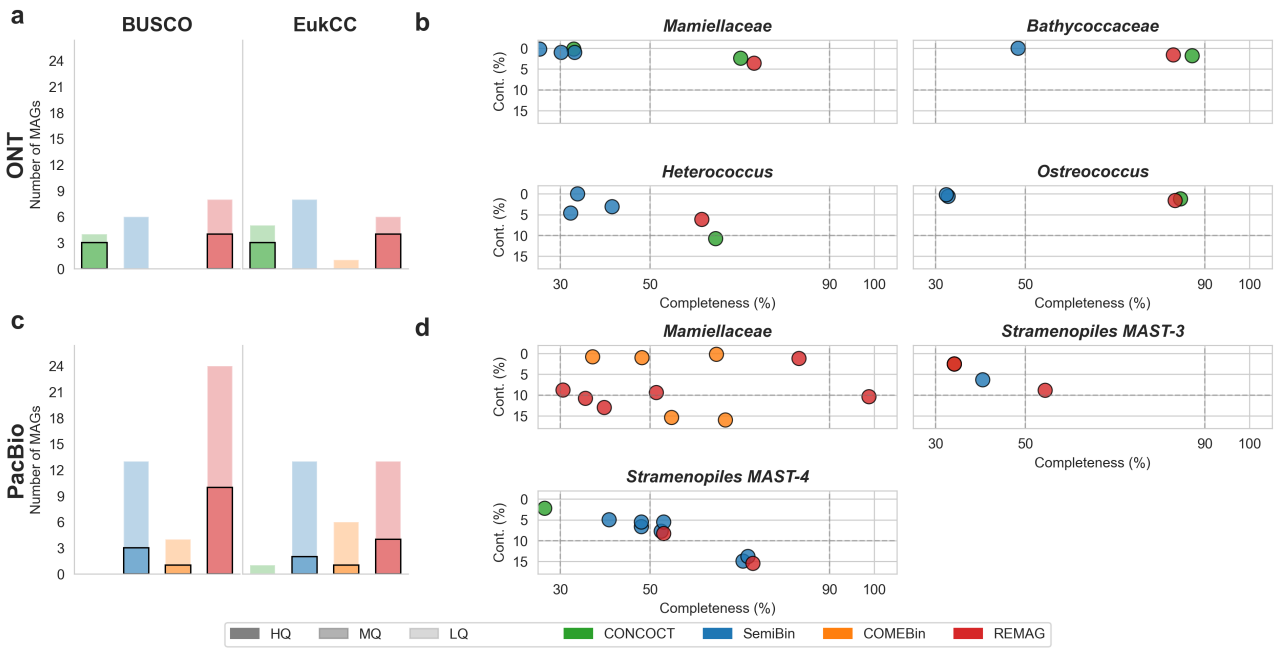

Figure 3. Eukaryotic MAG recovery performance on real plankton datasets
Comparison of binning tools (CONCOCT, SemiBin2, COMEBin and REMAG) on real plankton datasets sequenced with PacBio HiFi and Oxford Nanopore Technologies (ONT). **a** Shows the count of total MAGs with both BUSCO and EukCC for ONT and **b** the completeness/contamination according to EukCC for bins matched against taxonomic markers with at least a MAG classified as MQ (>50% completeness, <10% contamination). **c** and **d** show the results plots for the PacBio dataset.

been missed otherwise (Supplementary Figure S6). The rescue step had a large impact on the eARI which showed a substantial increase in the more diverse marine and soil datasets (marine: 0.42 to 0.79 [+88.5%], soil: 0.77 to 0.86 [+11.3%]).

Benchmark on short-read real data

We tested REMAG on short-read datasets from the *Tara* Oceans expedition at two locations and two different depths, surface and deep chlorophyll maximum (DCM) on fractions enriched for eukaryotes (20–180 µm; Supplementary Table S1). In both surface and DCM samples, REMAG consistently outperformed CONCOCT, SemiBin2, and COMEBin in the recovery of HQ and MQ eMAGs. In surface samples, REMAG recovered three MQ eMAGs according to BUSCO, compared to one by CONCOCT and none by the other tools. In the DCM, REMAG recovered two MQ eMAGs (BUSCO), while COMEBin recovered two and CONCOCT recovered none. Notably, according to EukCC, the only MQ eMAGs recovered across all stations and depths were exclusively identified by REMAG. While the overall recovery across all tools is limited, this likely reflects the significant divergence of these marine eukaryotic genomes from current reference databases and the inherent complexity of assembly and binning from complex marine communities using short-read sequencing.

Benchmark on long-read real data

To demonstrate REMAG's performance on long-read datasets, we tested REMAG on plankton sequenced with ONT and PacBio HiFi consisting of four samples each.

Long-read plankton datasets. REMAG recovered more than double the number of high-quality eMAGs when compared to other binners according to EukCC (8 vs. 3 for CONCOCT and 2 for SemiBin2). Although in the ONT dataset REMAG recovered a single additional MQ eMAG compared to CONCOCT, in the PacBio dataset, REMAG recovered four MQ eMAGs compared to two for SemiBin2 and none for CONCOCT (Figure 3a and c).

A closer look at the completeness and contamination of the recovered eMAGs shows a comparable performance for three out of the four MQ eMAGs in the ONT dataset between CONCOCT and REMAG, while for the Heterococcus taxon, the CONCOCT bin is slightly above the 10% contamination threshold (Figure 3b). In the PacBio HiFi dataset, comparing to the second-best performer SemiBin2, we see consistently higher performance for the MQ eMAGs in terms of completeness (e.g., 55% vs. 40% in SemiBin2 for MAST-3; Figure 3d). Although CONCOCT performs relatively well on the ONT dataset and SemiBin2 on the PacBio dataset, their performance is markedly inconsistent, as neither tool recovers any MQ MAGs in the other dataset. Meanwhile, REMAG reliably excels in both datasets. It is also noteworthy that although COMEBin ranks second in the simulated benchmarks, it is the worst-performing tool on long-read real datasets, recovering a single MQ eMAG across both datasets.

Plankton MAGs. We collated both plankton datasets (ONT and PacBio HiFi), including all recovered MQ and LQ eMAGs. We lowered the contamination cutoff to <15% due to the presence of high-completeness bins within this contamination range (Figure 4). Dereplicat-

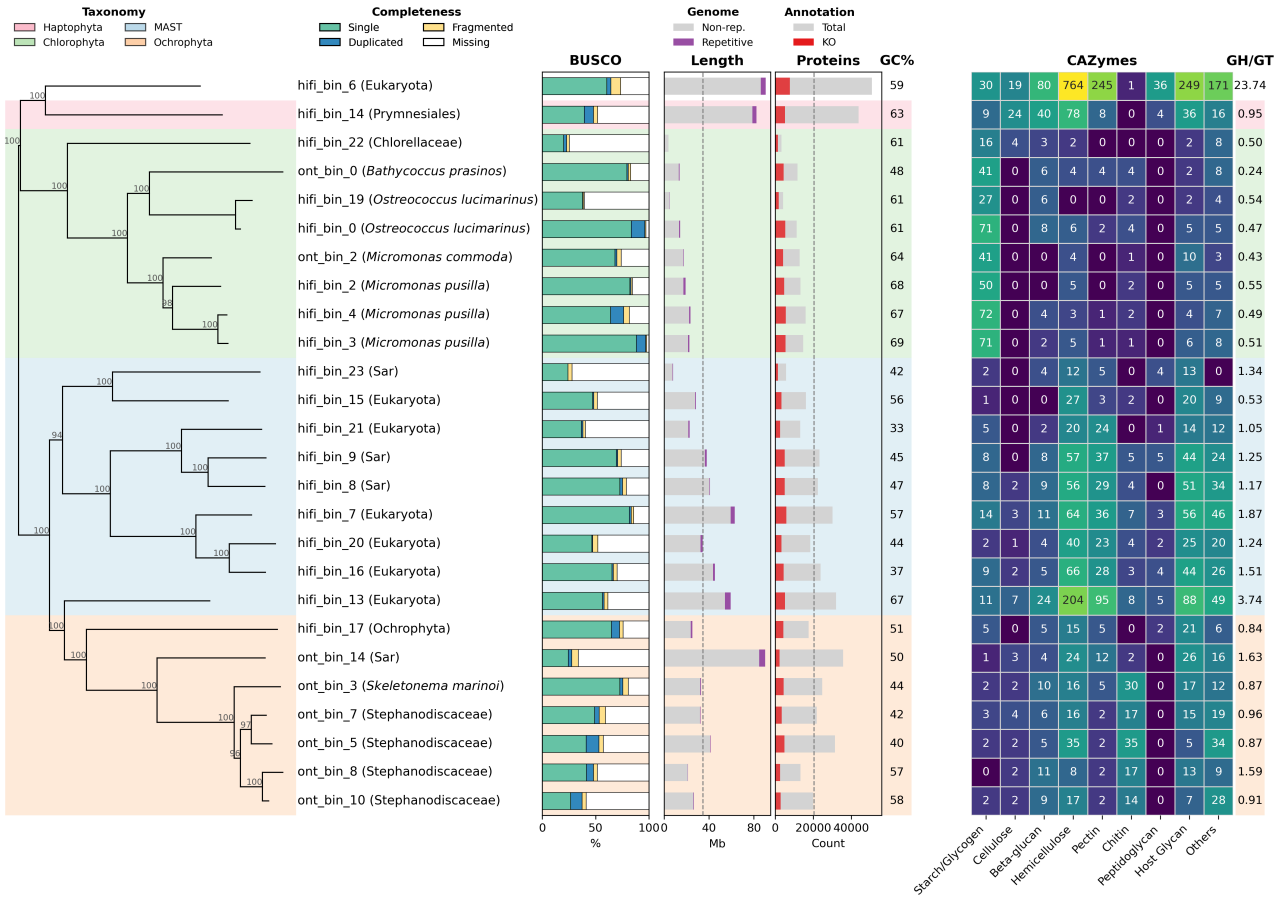

Figure 4. Eukaryotic MAGs recovered from plankton datasets
Phylogenetic tree of eukaryotic MAGs recovered from the plankton datasets sequenced with ONT and PacBio HiFi. The BUSCO results show the completeness of each MAG and in parentheses its taxonomy assignment. The length with annotated repetitive content and the number of predicted proteins with and without KO annotations is shown as stacked bar charts with the mean as a dashed line. Genome GC content is shown as a percentage. The heatmap indicates the number of carbohydrate-active enzymes (CAZymes) specific for different substrates and the ratio of glycoside hydrolases to glycosyltransferases (GH/GT) found in the MAGs.

ing the 38 bins at 99% average nucleotide identity (ANI) resulted in 26 unique taxonomically diverse eMAGs (24 at 95% ANI). These MAGs formed three distinct phylogenetic clusters: one composed of green algae with eight members, another of marine stramenopiles with 16 members (of which seven belonged to the phylum Ochrophyta), and a third cluster with two members including one haptophyte.

Most eMAGs showed a high completeness of core pathways (Supplementary Figure S8). However, there were variations which reflect both the difference in recovered completeness and the taxonomy, showing a clear difference between the two main clades of green algae and stramenopiles. For example, stramenopiles in the dataset had distinct secondary metabolism biosynthesis modules compared to the green algae, which reflects their evolutionary history and niche within marine ecosystems (Amsler and Fairhead, 2005).

We also noticed differences in the distribution of carbohydrate-active enzymes (CAZymes) among the eMAGs (Figure 4). When grouped by putative substrate type, there were patterns of abundance correlating to taxonomy and lifestyle. These included a larger number of starch/glycogen-related enzymes in green algae, which might relate to the storage of polysaccharides. For the stramenopiles, the higher abundance of CAZymes with substrates including hemicellulose, pectin, peptidoglycan and chitin, likely facilitates the degradation of complex carbohydrates in the marine environment, potentially including those derived from cell walls of other organisms (Jirsová and Wideman, 2024; Christaki et al., 2017). Additionally, we found a higher glycoside hydrolase to glycosyltransferase ratio in the stramenopiles, compared to the green algae, which reflects their higher heterotrophic potential (Seeleuthner et al., 2018).

Discussion

The recovery of MAGs is a critical step in understanding the composition and function of microbial communities. This is particularly true of eukaryotes as, despite their relevance for the environment, they are often underrepresented in databases due to challenges with lab culturing and difficulties in generating MAGs from metagenomes. In this study, we present REMAG, a tool designed to recover eMAGs from metagenomic samples, which we have benchmarked against existing

state-of-the-art binners. The use of contrastive learning and dedicated eukaryotic markers allows REMAG to effectively capture the complex features of eukaryotic genomes, leading to improved binning performance.

Compared to other contrastive-learning-based binners such as SemiBin2 and COMEBin, REMAG uses a Barlow Twins loss that requires only positive examples and no negative examples. This prevents noise caused by randomly sampled negative pairs that might inadvertently originate from the same genome and is particularly useful for the larger eukaryotic genomes which might be more fragmented. The exclusive use of positive links also reduces the large batch size requirements, allowing the same training volume to run faster on less memory. Furthermore, while COMEBin concatenates modalities for training and effectively weights the importance of coverage the same as composition, REMAG uses a fusion layer that adapts its weights to how informative each modality is by using cross-attention and gating mechanisms. This allows the training to dynamically adapt to the potentially variable coverage patterns of eukaryotic genomes by relying more on their composition depending on the dataset.

We have shown that REMAG outperforms other tools in recovering high-quality eMAGs from both simulated and real datasets. This is particularly evident when applied to long-read sequencing technologies, as shown by the broad phylogenetic range of highly complete eMAGs recovered from the plankton datasets. In short-read assemblies, however, REMAG's performance is comparable to that of CONCOCT (Alneberg et al., 2014). Despite being published over a decade ago, CONCOCT remains a strong performer for eukaryotic binning in short-read datasets. This is likely because CONCOCT does not rely on single-copy markers and lacks any assumptions about genome size. The reliance on accurate eukaryotic marker annotations for clustering is one of the limitations of REMAG, and is likely responsible for its modest performance on short-read datasets. The shorter reads result in shorter contigs on average, which might truncate protein-coding genes, resulting in less accurate marker annotation. While comparable in the simulated datasets, in real datasets as shown for the *Tara* Oceans samples, the performance was slightly better for REMAG, which demonstrates its utility in real-world applications where a large diversity of eukaryotes is expected.

While REMAG has shown improvement over competing tools, the low overall recovery of eukaryotes in the real datasets demonstrates the difficulty of this problem. This is compounded by the divergence of environmental eukaryotes from reference databases, which affects both quality assessment and marker-based clustering. Future efforts should be aimed at complementing coverage and composition with further intrinsic features that provide evidence for the same-genome origin of contigs, such as assembly graphs and Hi-C contact maps.

Conclusion

As sequencing technologies advance, the systematic study of the long and complex genomes of microbial eukaryotes from metagenomics is becoming increasingly feasible. In response to the demand for microbial eukaryote recovery tools from metagenomic data, we have developed REMAG. By leveraging contrastive learning and dedicated markers, it achieves superior high-quality eukaryotic MAG recovery compared to existing tools. This method is particularly valuable for exploring uncharacterized microbial diversity, as demonstrated for the plankton datasets. As metagenomic technical improvements continue and more samples become publicly available, tools like REMAG will be essential for comprehensive characterization of microbial communities across all domains of life.

Data and code availability

REMAG is open source and available at https://github.com/danielzmbp/remag under MIT license. The CAMI II plant-associated dataset used for benchmarking is available at https://data.cami-challenge.org/.

Methods

Dataset preparation

To evaluate REMAG's performance on realistic community structures, we generated synthetic metagenomic datasets representing three distinct biomes: marine, human gut, and soil. Taxonomic profiles and abundance distributions were derived from MGnify (Mitchell et al., 2020; Supplementary Table S2). The marine dataset (MGYS00001869) comprised 9 samples containing 428 organisms (370 prokaryotes and 58 eukaryotes). The human gut dataset (MGYS00002207) consisted of 19 samples with 478 organisms (464 prokaryotes and 14 eukaryotes). For the soil environment (MGYS00005779), we simulated 20 samples containing 291 organisms (244 prokaryotes and 47 eukaryotes). We downloaded the corresponding genomes from GenBank and RefSeq and generated random reads using RandomReadsMG (Bushnell et al., 2025) for short-read Illumina (scaled to 35 million reads), and long-read ONT (1 million reads) and PacBio (2.5 million reads) technologies. In addition, we evaluated REMAG's performance on the CAMI II (Meyer et al., 2022) plant-associated dataset for short reads and both ONT and PacBio using gold-standard pooled assemblies, the only CAMI dataset to include eukaryotes.
We assembled all real and simulated Illumina datasets using MEGAHIT (Li et al., 2015) and mapped them with BWA-MEM (Li, 2013). All real and simulated long-read datasets were assembled using metaMDBG v1.2 (Benoit et al., 2024) and mapped with minimap2 (Li, 2018).

REMAG pipeline

Data preprocessing. REMAG first filters for eukaryotic contigs using two classifiers based on HyenaDNA, a pre-trained genomic foundation model (Nguyen et al., 2023). Two tiny model variants with different context windows (1024 bp and 4096 bp) were fine-tuned on sequence chunks from 20,000 genomes from RefSeq (O'Leary et al., 2016): 10,000 prokaryotic (8,000 bacterial and 2,000 archaeal) and 10,000 eukaryotic (2,000 metazoan, 4,000 fungal and 4,000 protists). The fine-tuned HyenaDNA models process sequences using sliding windows corresponding to their context size (1024 bp model for contigs between 1000 and 4096 bp, and 4096 bp model for contigs >4096 bp) with adaptive stride lengths that scale with sequence length to optimize coverage: 256 bp for sequences <2 kb, 384 bp for sequences 2–3 kb, 512 bp for sequences 3–4 kb, 2048 bp for sequences 4–8 kb, 4096 bp for sequences 8–20 kb, and 8192 bp for sequences >20 kb. This adaptive approach balances computational efficiency with classification accuracy across the wide range of contig sizes typical in metagenomic assemblies and optimizes for increased recall for detection of eukaryotes. To improve classification confidence, the model employs early stopping when prediction confidence exceeds 90%, and applies random sampling of eight additional windows when confidence is below 75%. We retain contigs with eukaryotic probability scores $\geq$ 0.5 for downstream processing.

Data augmentation. REMAG employs a random masking strategy to generate diverse training views from each contig. For contigs exceeding 50 kb, sequences are first split at the midpoint. Four masking strategies chosen at random are applied: left, right, both edges, and central masking. A maximum of eight augmented fragments are generated per contig (or contig half) by default, with a maximum 25% overlap constraint between fragments to ensure diversity and reduce training on redundant features. All fragments must meet the minimum length threshold (default: 1000 bp for single-sample and 4096 bp for multi-sample). The process randomly creates training pairs where augmented views from the same contig and the original contig serve as positive examples (default: maximum 5,000,000 pairs). These positive pairs include all combinations of the original contig and augmented fragments.

Feature extraction. For each original contig and fragment, two feature modalities are computed: composition and abundance. Compositional features consist of canonical tetranucleotide (4-mer) frequencies (136 dimensions) calculated using reverse-complement aware counting, where each 4-mer and its reverse complement map to the same feature dimension (Teeling et al., 2004; Dick et al., 2009). The 136 dimensions represent the 256 possible 4-mers reduced by palindromes and reverse complement equivalence. Frequencies are normalized by fragment length with pseudocount addition (1e-5) to avoid zero-division.

Coverage features are calculated from alignment files (BAM/CRAM). For alignment data, per-base coverage is computed using pysam, a wrapper for SAMtools (Li et al., 2009), then averaged across fragment coordinates. Both mean coverage and standard deviation are computed per contig or fragment per sample. These raw coverage values are first normalized by the total mapped reads per sample (converting to reads per million) to account for differing sequencing depths. All normalized coverage values then undergo log1p transformation followed by global MinMax scaling to [0,1] range, preserving co-abundance patterns across samples.

Modal encoders. The network architecture consists of a dual-encoder Siamese network with separate multi-layer perceptrons for 4-mer frequencies (composition) and coverage (abundance) features. The composition encoder uses two fully connected layers (136→256→128) with batch normalization, LeakyReLU activation, and 5% dropout. The abundance encoder is adaptively sized based on sample count: for $\leq$2 samples (32→16 dimensions), 3–5 samples (64→32), 6–10 samples (128→64), and >10 samples (256→128), reflecting increasing co-abundance pattern complexity and the need for higher-dimensional representation.

Fusion layer. A fusion layer projects both modalities to a shared embedding space (default: 256 dimensions) using a bidirectional cross-attention mechanism (Vaswani et al., 2017; Chen et al., 2021). The fusion architecture employs multi-head attention modules (4 heads, 10% dropout) enabling both 4-mer and coverage features to attend to each other, capturing complementary compositional and abundance information.

To integrate these signals, we implement a gated fusion mechanism that computes a coefficient vector α to dynamically balance intrinsic features (h) against the retrieved cross-modal context (a). This operates via

$$z = \alpha \odot h + (1 - \alpha) \odot a$$

, allowing the model to selectively suppress noisy features by prioritizing the complementary modality. This integration is stabilized by a learnable residual weight w, clamped to the interval $[0.1, 0.9]$ as $\rho = \min(\max(w, 0.1), 0.9)$, and applied as $z' = (1 - \rho)z + \rho h$.

To model patterns at different granularities, we apply a multi-scale feature module to both gated modality vectors. The module uses three parallel transformations: a bottleneck path ($E \to \lfloor E/2 \rfloor \to E$), a same-scale linear path ($E \to E$), and an expansion path ($E \to 2E \to E$), where E is the embedding dimension. At each scale, transformed k-mer and coverage vectors are concatenated, and all scale-specific outputs are concatenated into a single multi-scale representation. This representation is then compressed and passed through a non-linear interaction block, followed by adaptive dropout.

Finally, the interaction output is concatenated with the gated modality vectors and the cross-attention alignment signal before the final fusion MLP. For datasets lacking coverage data, 4-mer features bypass the cross-attention stage and are processed through a simple projection layer.

Projection head and training. A projection head (256→512→512 dimensions) with batch normalization and ReLU activation transforms embeddings for loss calculation during training. The network is trained exclusively on positive pairs using Barlow Twins loss (Zbontar et al., 2021) (default: $\lambda = 3 \times 10^{-3}$ for single-sample and $\lambda = 2 \times 10^{-2}$ for multi-sample), which encourages invariance across augmented views while reducing redundancy across embedding dimensions through cross-correlation matrix regularization. The loss function is defined as:

$$\mathcal{L}_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2$$

where $\mathcal{C}$ is the cross-correlation matrix of the batch-normalized embeddings from the two identical networks where $\mathcal{C} \in \mathbb{R}^{d \times d}$ is the cross-correlation matrix computed across the batch dimension, with d being the projection head output dimensionality. The first term minimizes the difference between diagonal elements and 1 (invariance objective), while the second term minimizes off-diagonal elements (redundancy reduction objective). Optimization employs AdamW (Loshchilov and Hutter, 2017) (β_1=0.9, β_2=0.95, weight decay=0.05) with gradient clipping (max norm=1.0). Learning rate scheduling includes a 10-epoch linear warmup from 10% to full learning rate, followed by cosine annealing to 1% of peak rate. The base learning rate defaults to 5×10^{-3} for single-sample datasets (or 5×10^{-4} for co-assemblies), and is linearly scaled by $(batch_size/256) \times 0.2$. Training uses a default batch size of 2048 for 400 epochs with early stopping monitoring training loss (patience=20 epochs).

Graph construction and marker annotation. A k-NN graph is constructed from L2-normalized embeddings resulting from the fusion layer (default: 256 dimensions) using cosine similarity. The default parameters include k=15 neighbors per node and a similarity threshold of 0.1 to prune weak edges. Graph construction is parallelized using sklearn's NearestNeighbors (Pedregosa et al., 2011) with brute-force search. The resulting directed graph is converted to an undirected graph by averaging reciprocal edge weights. All contigs are annotated by mapping the 133 families of eukaryotic single-copy core genes (SCGs) from ORTHODB version 12 (Kuznetsov et al., 2023) using Miniprot (Li, 2023).

Greedy iterative clustering. REMAG employs a greedy iterative clustering strategy to identify high-quality bins while avoiding the limitations of a single global resolution parameter. The algorithm iteratively extracts bins from the k-NN graph using the following process. (1) The current graph (initially containing all contigs) is clustered using the Leiden algorithm (Traag et al., 2019) across a range of resolutions (default: 0.1, 0.5, 1.0, 2.0, 5.0). (2) For every cluster generated at each resolution, a quality score is calculated based on SCGs. The score utilizes an F1 metric, defined as the harmonic mean of completeness ($C = N_{\text{unique}}/133$) and precision ($P = 1 - K$), where K is contamination defined as the fraction of total marker genes found (G) that are redundant ($K = \frac{G - N_{\text{unique}}}{G}$):

$$F1 = 2 \cdot \frac{C \cdot P}{C + P}$$

. (3) Before selection, candidate clusters are filtered using a strict duplication threshold; clusters where the number of duplicated single-copy core gene families exceeds 10% of the total 133 marker families ($\frac{N_{\text{duplicated}}}{133} > 0.10$) are rejected. The valid cluster with the highest F1 score across all tested resolutions is then selected as a bin, and the contigs belonging to this bin are removed from the graph. (4) The process repeats with the remaining nodes until no further valid bins can be formed.

Satellite rescue. Following clustering, REMAG employs a strategy to re-merge bins that might have been over-fragmented during the initial clustering step. This step aims to merge smaller "satellite" bins into larger "core" bins based on their embedding similarity, subject to contamination constraints. For each bin, a weighted centroid is calculated in the high-dimensional embedding space, where each contig's embedding vector is weighted by its sequence length. Bins are then sorted by total size (in base pairs). Iterating from smallest to largest, each bin is treated as a potential candidate for merging. The algorithm searches for a target bin among larger bins that shares the highest cosine similarity with the candidate's centroid, provided the similarity exceeds a defined threshold (default: 0.70). To ensure genome quality is preserved, a safety check is performed before any merge. The SCG duplication rate ($\frac{N_{duplicated}}{N_{unique}}$) is calculated for the hypothetical merged bin. The merge is only executed if the absolute increase in the SCG duplication rate does not exceed 10%, and the total duplication rate of the merged bin remains $\leq$10%, ensuring that distinct genomes are not erroneously combined.

Recovery quality assessment

Following established standards for metagenomics, high quality (HQ) MAGs were defined as having >90% completeness and <10% contamination (measured by SCG duplication as in the K clustering metric). Medium quality (MQ) MAGs were defined as having >50% completeness and <10% contamination. Bins with >30% and <50% completeness and <10% contamination were classified as low quality (LQ). For datasets with ground truth, we mapped the binned contigs back to

the reference genomes using minimap2 (Li, 2018) and generated a contig-to-genome mapping file for each dataset. For real datasets, we used BUSCO (Manni et al., 2021) with the eukaryota_odb12 database and EukCC (Saary et al., 2020) to estimate completeness and contamination based on single-copy core gene presence and duplication.

Clustering accuracy was quantified using the Adjusted Rand Index (ARI) computed from ground-truth contig-to-genome assignments and predicted bin labels. We report a eukaryote ARI (eARI) weighted by contig length. This eARI includes only bins that contain at least one eukaryotic contig, plus unbinned eukaryotic contigs, to penalize bins that mix eukaryotic and non-eukaryotic content.

Plankton MAG downstream analysis

For downstream characterization of plankton MAGs, we generated a phylogeny from single-copy marker proteins annotated with BUSCO (Manni et al., 2021) and the eukaryota_odb10 lineage present in more than 50% of MAGs, due to the larger pool of markers compared to ODB12. Marker protein sequences were aligned with MAFFT (Katoh and Standley, 2013), and a phylogenetic tree was inferred with IQ-TREE v3 (Wong et al., 2025). Gene calling was performed with BRAKER3 (Gabriel et al., 2024) on the repeat-masked (Smit et al., 2013–2015) contigs. Taxonomy assignment was performed using MMseqs2 (Steinegger and Söding, 2017) against Uniprot90 (The UniProt Consortium, 2023). Functional annotation was performed by assigning KEGG Orthologs (KOs) with KofamScan using profile HMMs and adaptive score thresholds (Aramaki et al., 2020).

Computational resources

All analyses were executed on a single node equipped with 32 CPU cores, 128 GB RAM and an NVIDIA A100 GPU with 80 GB of memory.

Acknowledgements

This work was supported by the BBSRC (UKRI) through the Earlham Institute Strategic Programme Grant BBX011089/1 (WP1: BBS/E/ER/230002A) and Core Capability Grant BB/CCG2220/1. We thank the Earlham Institute Research Computing Group for providing High Performance Computing resources.

Bibliography

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146. doi:10.1038/nmeth.3103.

Amsler, C. D. and Fairhead, V. A. (2005). Defensive and sensory chemical ecology of brown algae. *Advances in Botanical Research*, 43(Ecological Monographs752005):1–91. doi:10.1016/s0065-2296(05)43001-3.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). Kofamkoala: Kegg ortholog assignment based on profile hmm and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252. doi:10.1093/bioinformatics/btz859.

Benoit, G., Raguideau, S., James, R., Phillippy, A. M., Chikhi, R., and Quince, C. (2024). High-quality metagenome assembly from long accurate reads with metamdbg. *Nature Biotechnology*. doi:10.1038/s41587-023-01983-6.

Bushnell, B., Schulz, F., and Villada, J. C. (2025). Rapid terabase-scale simulation of realistic metagenomes for experimental design and pathogen detection with RandomReadsMG. *bioRxiv*, page 2025.07.18.665570. doi:10.1101/2025.07.18.665570.

Chen, C.-F., Fan, Q., and Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356, (2021).

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, (2020). doi:10.48550/arXiv.2002.05709.

Christaki, U., Genitsaris, S., Monchy, S., Li, L. L., Rachik, S., Breton, E., and Sime-Ngando, T. (2017). Parasitic eukaryotes in a meso-eutrophic coastal system with marked phaeocystis globosa blooms. *Frontiers in Marine Science*, 4:416. doi:10.3389/fmars.2017.00416.

Cuscó, A., Duan, Y., Gil, F., Chklovski, A., Kruthi, N., Pan, S., Forslund, S., Lau, S., Löber, U., Zhao, X.-M., and Coelho, L. P. (2025). Capturing global pet dog gut microbial diversity and hundreds of near-finished bacterial genomes by using long-read metagenomics in a shanghai cohort. *bioRxiv*, page 2025.09.17.676595. doi:10.1101/2025.09.17.676595.

del Campo, J., Sieracki, M. E., Molestina, R., Keeling, P., Massana, R., and Ruiz-Trillo, I. (2014). The others: our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, 29(5):252–259. doi:10.1016/j.tree.2014.03.006.

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A. M., Kourlaiev, A., d'Agata, L., Clayssen, Q., et al. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Systems*, 33(5):364–379. doi:10.1016/j.cels.2022.04.006.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., and Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, 10(8):R85. doi:10.1186/gb-2009-10-8-r85.

Elliott, T. A. and Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140331. doi:10.1098/rstb.2014.0331.

Eraslan, G., Avsec, Z., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7):389–403. doi:10.1038/s41576-019-0122-6.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science*, 320(5879):1034–1039. doi:10.1126/science.1153253.

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., and Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*, 34(5):769–777. doi:10.1101/gr.278090.123.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351. doi:10.1038/nrg.2016.49.

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239. doi:10.1186/s13059-016-1103-0.

Jirsová, D. and Wideman, J. G. (2024). Integrated overview of stramenopile ecology, taxonomy, and heterotrophic origin. *The ISME Journal*, 18(1):wrae150. doi:10.1093/ismejo/wrae150.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359. doi:10.7717/peerj.7359.

Karlicki, M., Antonowicz, S., and Karnkowska, A. (2022). Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics*, 38(2):344–350. doi:10.1093/bioinformatics/btab672.

Katoh, K. and Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780. doi:10.1093/molbev/mst010.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, (2020). doi:10.48550/arXiv.2004.11362.

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., and Zdobnov, E. M. (2023). Orthodb v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, 51(D1):D445–D451. doi:10.1093/nar/gkac998.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. doi:10.1038/nature14539.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676. doi:10.1093/bioinformatics/btv033.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*. doi:10.48550/arXiv.1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li, H. (2023). Miniprot: fast mapping of proteins to genomes. *Bioinformatics*, 39(1). doi:10.1093/bioinformatics/btad014.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Loshchilov, I. and Hutter, F. (2017). Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*. doi:10.48550/arXiv.1711.05101.

Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science*, 302(5649):1401–1404. doi:10.1126/science.1089370.

Manni, M., Berkeley, M. R., Seppey, M., and Zdobnov, E. M. (2021). Busco update: novel

and streamlined workflows for quality assessments, phylogenomics, and gene prediction in eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10): 4647–4654.

Massana, R. (2011). Eukaryotic picoplankton in surface oceans. *Annual Review of Microbiology*, 65:91–110. doi:10.1146/annurev-micro-090110-102903.

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., et al. (2022). Critical assessment of metagenome interpretation: the second round of challenges. *Nature Methods*, 19(4):429–440. doi:10.1038/s41592-022-01431-4.

Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., et al. (2020). Mgnify: the microbiome analysis resource in 2020. *Nucleic acids research*, 48(D1):D570–D578. doi:10.1093/nar/gkz1035.

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Birch-Sykes, C., Wornow, M., Patel, A., Rabideau, C., Massaroli, S., Bengio, Y., Ermon, S., Ré, C., and Baccus, S. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In *Advances in Neural Information Processing Systems*, (2023). doi:10.48550/arXiv.2306.15794.

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., and Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, 39(5):555–560. doi:10.1038/s41587-020-00777-4.

O'Leary, N. A., Wright, M. W., Brister, J., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745. doi:10.1093/nar/gkv1189.

Pan, S., Zhu, C., Zhao, X.-M., and Coelho, L. P. (2023). Semibin2: self-supervised contrastive learning leads to better mags for short- and long-read sequencing. *Bioinformatics*, 39 (Supplement_1):i21–i29. doi:10.1093/bioinformatics/btad209.

Parfrey, L. W., Walters, W. A., and Knight, R. (2011). Microbial eukaryotes in the human microbiome: Ecology, evolution, and future directions. *Frontiers in Microbiology*, 2:153. doi:10.3389/fmicb.2011.00153.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pu, L. and Shamir, R. (2024). 4cac: 4-class classifier of metagenome contigs using machine learning and assembly graphs. *Nucleic Acids Research*, 52(19):e94. doi:10.1093/nar/gkae799.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9):833–844. doi:10.1038/nbt.3935.

Saary, P., Mitchell, A. L., and Finn, R. D. (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with eukcc. *Genome Biology*, 21(1): 244. doi:10.1186/s13059-020-02155-4.

Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., Leconte, J., Mangot, J.-F., Poulain, J., Labadie, K., Logares, R., Sunagawa, S., Berardinis, V. d., Salanoubat, M., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Coordinators, T. O., Acinas, S. G., Boss, E., Follows, M., Gorsky, G., Grimsley, N., Karp-Boss, L., Krzic, U., Not, F., Ogata, H., Raes, J., Reynaud, E. G., Sardet, C., Speich, S., Stemmann, L., Velayoudon, D., Weissenbach, J., Pesant, S., Poulton, N., Stepanauskas, R., Bork, P., Bowler, C., Hingamp, P., Sullivan, M. B., Iudicone, D., Massana, R., Aury, J.-M., Henrissat, B., Karsenti, E., Jaillon, O., Sieracki, M., Vargas, C. d., and Wincker, P. (2018). Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nature Communications*, 9(1):310. doi:10.1038/s41467-017-02235-3.

Singleton, C. M., Jensen, T. B. N., Delogu, F., Knudsen, K. S., Sørensen, E. A., Jørgensen, V. R., Karst, S. M., Yang, Y., Sereika, M., Petriglieri, F., Knutsson, S., Dall, S. M., Kirkegaard, R. H., Kristensen, J. M., Overgaard, C. K., Woodcroft, B. J., Speth, D. R., Aroney, S. T. N., Consortium, T. M. D., Thomsen, H. C., Christensen, B. T., Jonge, L. W. d., Danielsen, A.-C. S., Hermansen, C., Greve, M. H., Ejrnæs, R., Davidson, T. A., Normand, S., Treier, U. A., Madsen, B., Schramm, A., Marshall, I. P. G., Dam, A.-S., Kjeldsen, K. U., Finster, K., Thomsen, P. F., Sigsgaard, E. E., Klepke, M. J., Vestergård, M., Aude, E., Thomsen, L., Lemming, C., Hørfarter, R., Jensen, M. M., Frøslev, T. G., Gram, L., Svendsen, P. B., Schostag, M. D., Kjellerup, S., Skovhus, T. L., Søborg, D. A., Reitzel, K., Pedersen, J. F., Giguere, A., Pedersen, I. S., Sønderkær, M., Vollertsen, J., Liu, F., Roslev, P., Iversen, N., Nielsen, K. L., Jonge, N. d., Bruhn, D., Nielsen, J. L., Kristensen, T. N., Jiang, C., Nierychlo, M. A., Dottorini, G., Wagner, M., Dueholm, M. K. D., Nielsen, P. H., and Albertsen, M. (2026). The Microflora Danica atlas of Danish environmental microbiomes. *Nature*, 649(8098):971–981. doi:10.1038/s41586-025-09794-2.

Smit, A. F. A., Hubley, R., and Green, P. Repeatmasker open-4.0, (2013–2015). URL http://www.repeatmasker.org.

Steinegger, M. and Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028. doi:10.1038/nbt.3988.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359. doi:10.1126/science.1261359.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. O. (2004). Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:163. doi:10.1186/1471-2105-5-163.

The UniProt Consortium. (2023). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531. doi:10.1093/nar/gkac1052.

Traag, V. A., Waltman, L., and van Eck, N. J. (2019). From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233. doi:10.1038/s41598-019-41695-z.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, (2017). doi:10.48550/arXiv.1706.03762.

Wang, Z., You, R., Han, H., Liu, W., Sun, F., and Zhu, S. (2024). Comebin: effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nature Communications*, 15:575. doi:10.1038/s41467-023-44290-z.

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., and Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, 28(4):569–580. doi:10.1101/gr.228429.117.

Wong, T., Ly-Trong, N., Ren, H., Baños, H., Roger, A., Susko, E., Bielow, C., Maio, N. D., Goldman, N., Hahn, M., Huttley, G., Lanfear, R., and Minh, B. Q. (2025). IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. doi:10.32942/x2p62n.

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2:26. doi:10.1186/2049-2618-2-26.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320, (2021). doi:10.48550/arXiv.2103.03230.

Zhao, D., Salas-Leiva, D. E., Williams, S. K., Dunn, K. A., Shao, J. D., and Roger, A. J. (2025). Eukfinder: a pipeline to retrieve microbial eukaryote genome sequences from metagenomic data. *mBio*, 16(5):e00699–25. doi:10.1128/mbio.00699-25.

Supplementary Information

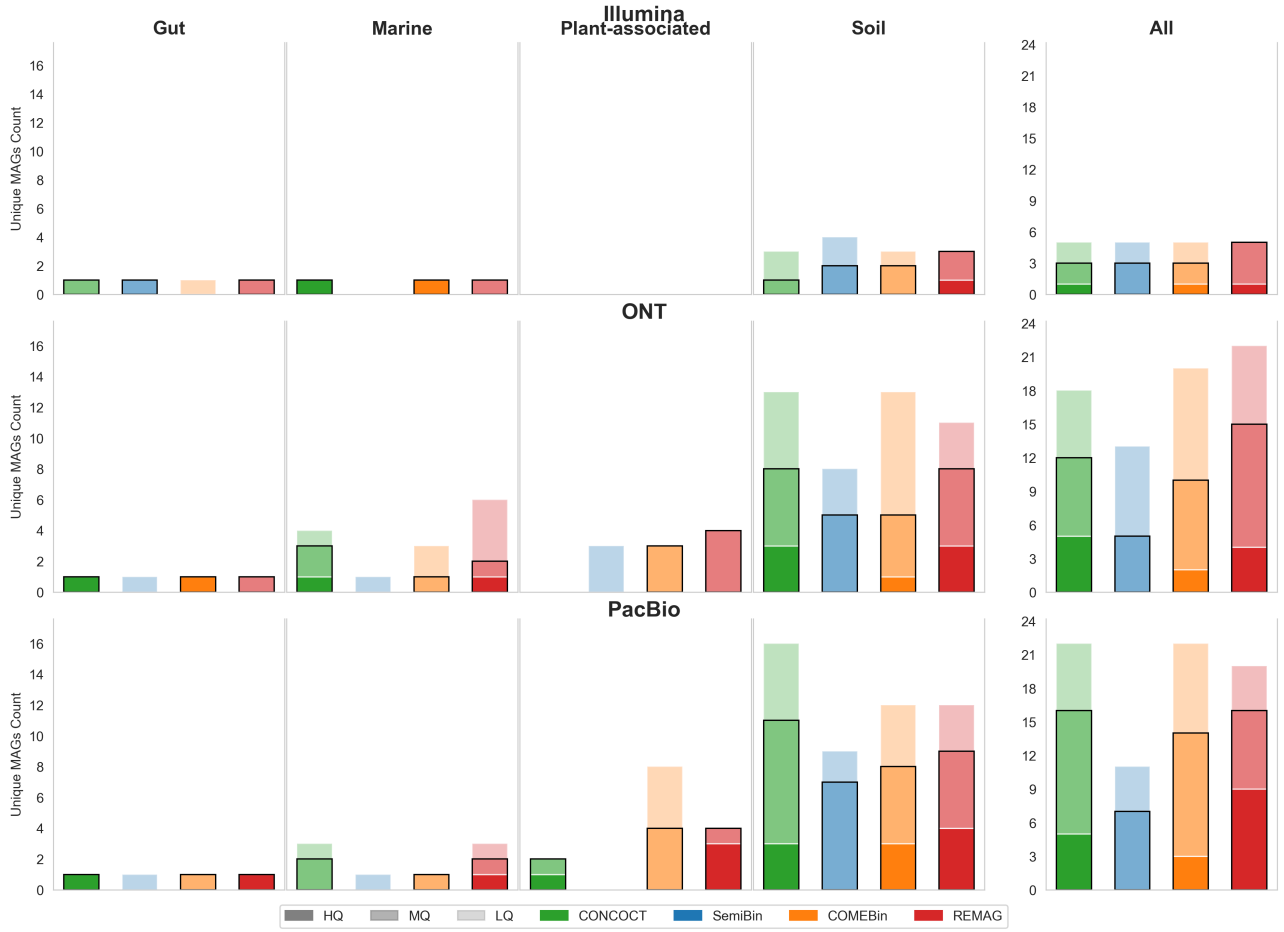

Figure S1. MAG recovery performance in per-sample assemblies
Shown are the number of unique MAGs recovered per sample for CONCOCT, SemiBin2, COMEBin and REMAG across all simulated datasets for each of the three sequencing technologies.

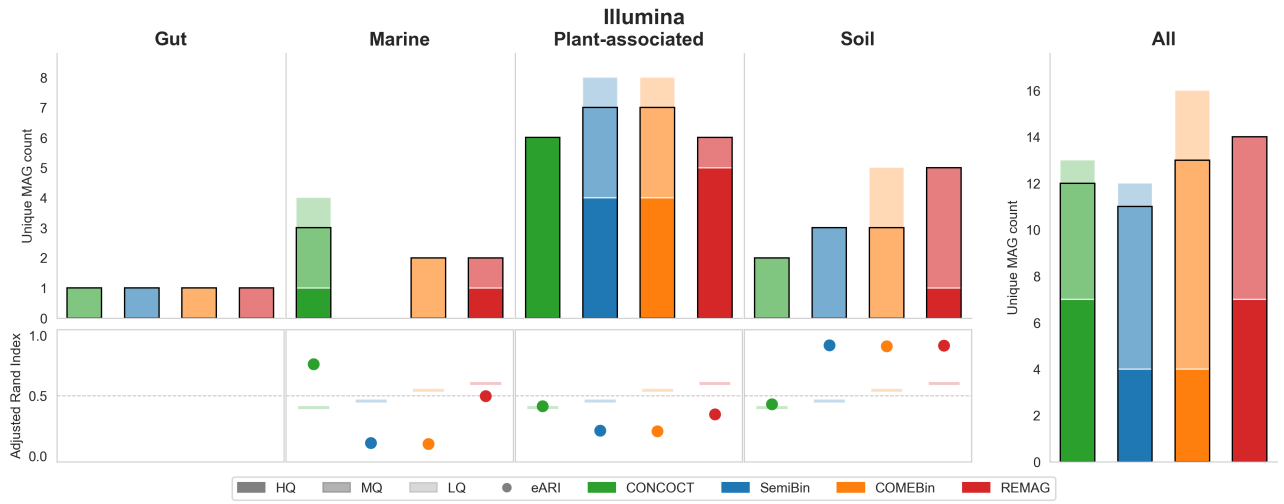

Figure S2. Short-read simulated dataset eukaryotic MAG recovery
Comparison of short-read MAG recovery statistics for CONCOCT, SemiBin2, COMEBin and REMAG across all four simulated datasets. Note, the Adjusted Rand Index (ARI) is not shown for the Gut dataset as it only contains one recoverable eukaryotic genome.

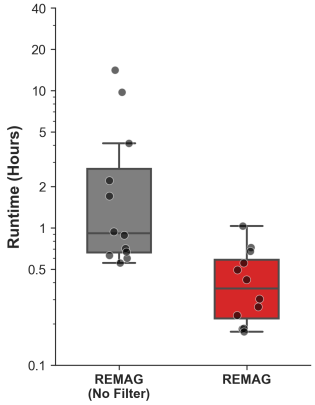

Figure S3. Impact of eukaryotic contig filtering on REMAG runtime
Comparison of REMAG execution times with and without the eukaryotic contig filtering step across all simulated datasets.

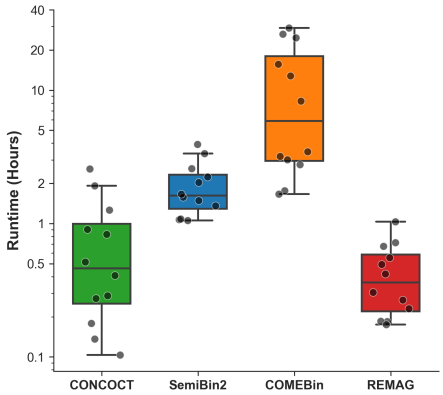

Figure S4. Runtime comparison across binning tools.
Comparison of execution times for CONCOCT, SemiBin2, COMEBin, and REMAG across all simulated datasets for the different sequencing technologies.

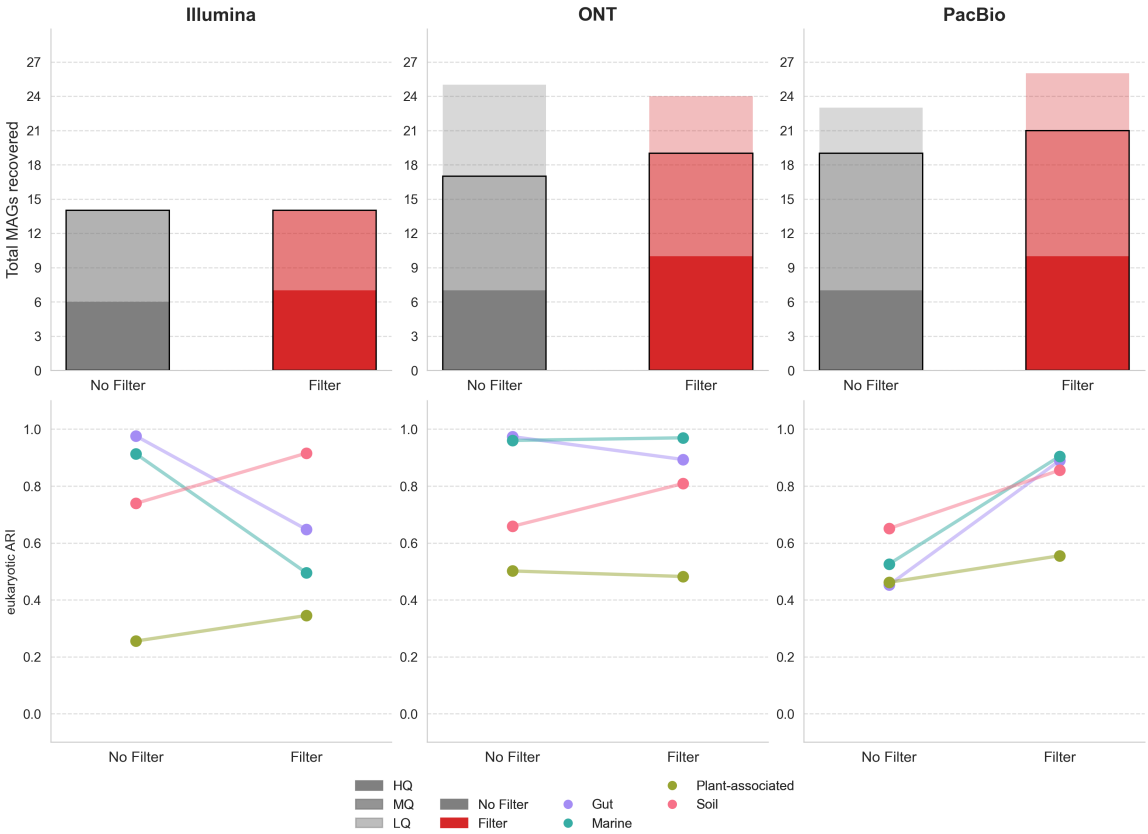

Figure S5. Ablation testing of eukaryotic contig filtering step
Comparison of REMAG with and without the eukaryotic contig filtering step on simulated gut, soil, marine and plant-associated datasets for the recovery of eukaryotic MAGs.

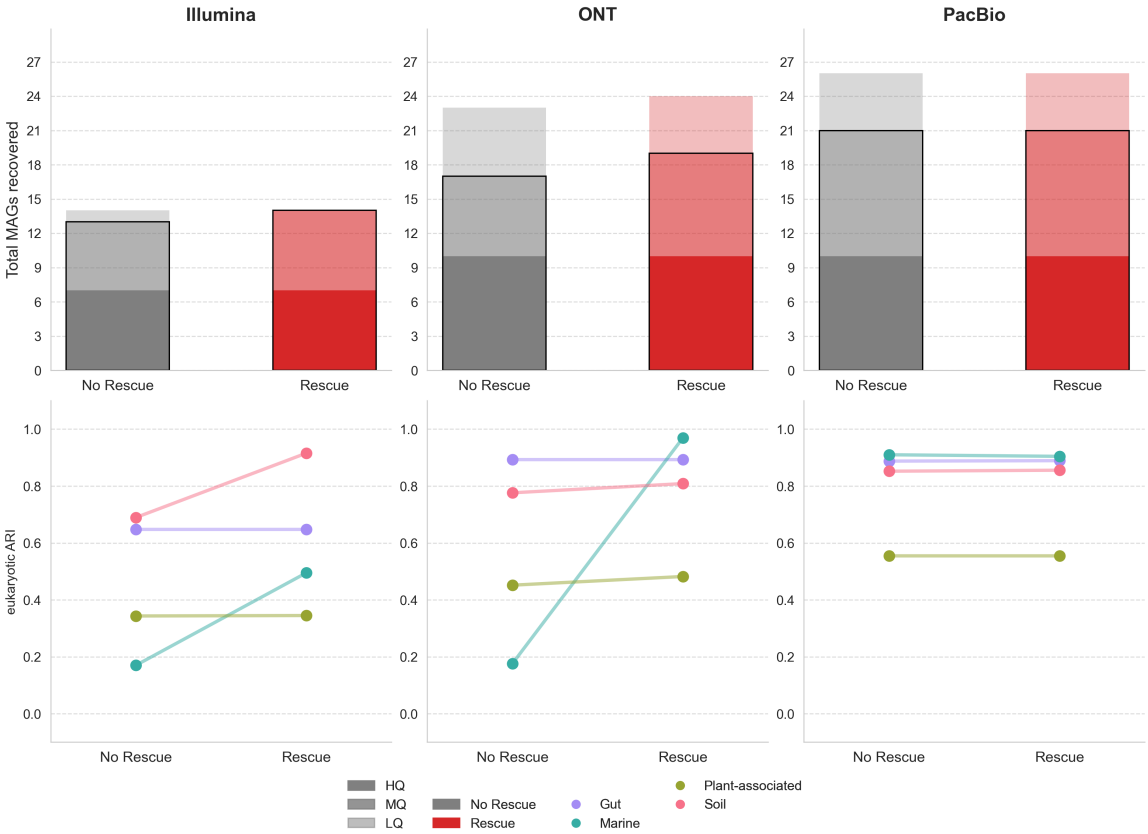

Figure S6. Ablation testing of satellite rescue step
Comparison of REMAG performance with and without the satellite rescue step on simulated gut, soil, marine and plant-associated datasets for the recovery of eukaryotic MAGs.

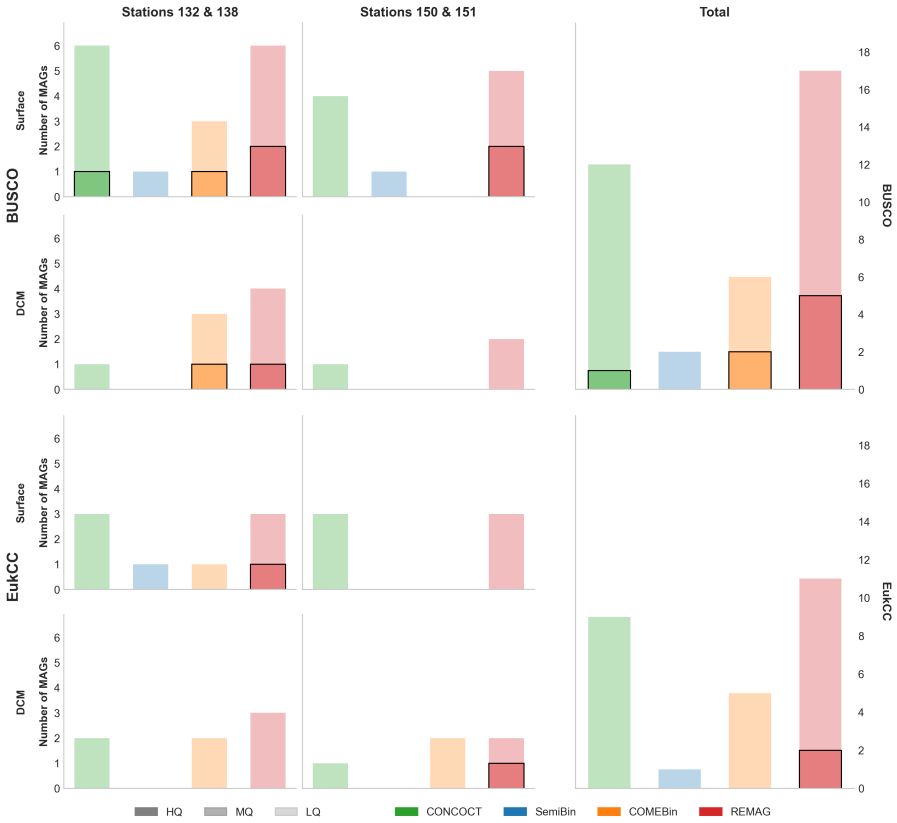

Figure S7. Eukaryotic MAG recovery performance on short-read *Tara* Oceans datasets
Comparison of binning tools (CONCOCT, SemiBin2, COMEBin and REMAG) on *Tara* Oceans short-read co-assemblies from two locations (station 132 & 138) and (station 150 & 151) at two depths, surface and deep chlorophyll maximum (DCM). Bar plots show the number of MAGs at different quality tiers assessed by BUSCO (eukaryota_odb12) and EukCC.

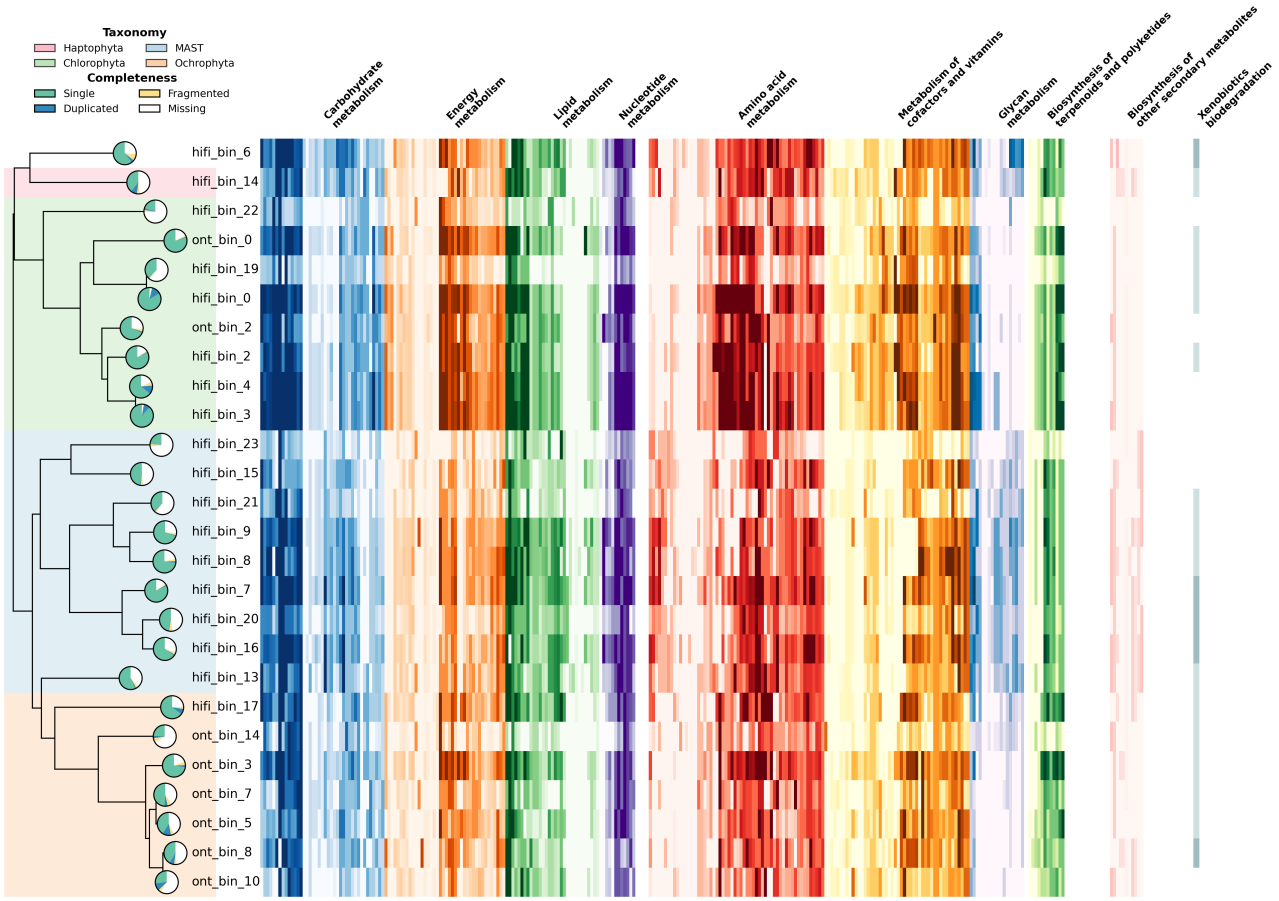

Figure S8. KEGG pathway completeness for plankton MAGs
Heatmap showing the completeness of all KEGG modules identified in the recovered eukaryotic MAGs from the plankton datasets.

Supplementary Tables

Sampling zone	Station	Run accession
Surface	132	ERR1726872
Surface	132	ERR1726580
Surface	138	ERR1726967
Surface	138	ERR1726948
DCM	132	ERR1726769
DCM	138	ERR1726874
Surface	150	ERR1726860
Surface	151	ERR1726744
DCM	150	ERR1726857
DCM	151	ERR1726743

Table S1. *Tara* Oceans sequencing runs used for short-read benchmarking.
Paired-end Illumina reads from four *Tara* Oceans stations in the North Atlantic at surface and deep chlorophyll maximum (DCM), for the 20–180 µm size fraction enriched for eukaryotes. All runs are available from the European Nucleotide Archive (ENA).

Biome	MGnify Study	Sample Accessions
Marine	MGYS00001869	ERS1800270–ERS1800278
Human Gut	MGYS00002207	ERS2181189–ERS2181198, ERS2181200–ERS2181208
Soil	MGYS00005779	ERS6356179–ERS6356181, ERS6356183–ERS6356184, ERS6356186, ERS6356192, ERS6356197, ERS6356203, ERS6356207, ERS6356214, ERS6356218–ERS6356219, ERS6356222, ERS6356224–ERS6356225, ERS6356231, ERS6356237–ERS6356238

Table S2. MGnify study accessions used for the simulated datasets.
The three biomes (marine, human gut, and soil) and their corresponding MGnify study and sample accessions used for generating taxonomic profiles and abundance distributions for the simulated datasets.