

RESOURCE ARTICLE OPEN ACCESS

ChloroScan: Recovering Plastid Genome Bins From Metagenomic Data

Yuhao Tong^{1,2}  | Vanessa Rossetto Marcelino^{3,4} | Robert Turnbull⁵  | Heroen Verbruggen⁶

¹School of Computing and Information Systems, The University of Melbourne, Carlton, Victoria, Australia | ²Melbourne Integrative Genomics, School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia | ³Institute of Agrochemistry and Food Technology, Spanish National Research Council (CSIC), Valencia, Spain | ⁴Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, Victoria, Australia | ⁵Melbourne Data Analytics Platform (MDAP), The University of Melbourne, Melbourne, Victoria, Australia | ⁶CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto, Vairão, Portugal

Correspondence: Yuhao Tong (yuhao.tong@student.unimelb.edu.au)**Received:** 9 October 2025 | **Revised:** 23 March 2026 | **Accepted:** 2 April 2026**Keywords:** algae | bioinformatics | genome-resolved metagenomics | microbiome | plastid

ABSTRACT

Genome-resolved metagenomics has contributed greatly to discovering prokaryotic genomes. When applied to microscopic eukaryotes (protists), challenges such as the high number of introns and repeat regions found in nuclear genomes have hampered the mining and discovery of novel protistan lineages. Organellar genomes are simpler, smaller, have higher abundance than their nuclear counterparts and contain valuable phylogenetic information, but are yet to be widely used to identify new protist lineages from metagenomes. Here we present “ChloroScan”, a new bioinformatics pipeline to extract eukaryotic plastid genomes from metagenomes. It incorporates a deep learning contig classifier to identify putative plastid contigs and an automated binning module to recover bins with guidance from a curated marker gene database. Additionally, ChloroScan summarizes the results in different user-friendly formats, including annotated coding sequences and proteins for each bin. We show that ChloroScan recovers more high-quality plastid bins than MetaBAT2 for simulated metagenomes. The practical utility of ChloroScan is illustrated by recovering 16 medium to high-quality metagenome assembled genomes (MAGs) from four protist-size-fraction metagenomes, with several bins showing high taxonomic novelty. The ChloroScan code (v0.1.7) is available at https://github.com/Andyarguea/sae/chloroscan/tree/release_v0.1.7 under Apache-2.0 licence.

1 | Introduction

The amount of sequenced data for microbiomes has ballooned in the last decade (Quince et al. 2017). Genome-resolved metagenomics (GRM) became a widely used approach to analyse environmental microbiome data, offering numerous insights into their evolution, ecology and diversity (Salazar et al. 2022). It incorporates *de novo* assembly to build contigs and binning algorithms to cluster these into metagenome-assembled genomes (MAGs). These MAGs can then be used in phylogenetic analyses (Nayfach et al. 2021), or to understand microbial metabolism, adaptations to different environments (Xiao et al. 2025), or to

gain insights into the genomics of biological interactions (Dai et al. 2025; Yang et al. 2022).

GRM has greatly expanded our knowledge of prokaryotes. For example, it has led to the discovery of nanosized Candidate Phyla Radiation (CPR) representing at least 15% of bacterial phyla (Brown et al. 2015), and the discovery of Asgard Archaea that drastically altered our understanding of eukaryote origins (Hug et al. 2016; Liu et al. 2021; Leão et al. 2024). However, metagenomics has not yet achieved similar breakthroughs for microbial eukaryotes (protists) (Eme and Tamarit 2024). Like prokaryotes, microbial eukaryotes have significant

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2026 The Author(s). *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

roles in different environments (Chabé et al. 2017; Diaz and Plummer 2018; Solomon et al. 2022), and their genomes are valuable resources to understand their physiology, evolution, and ecological niche (Sibbald and Archibald 2017).

Currently, only a minor fraction of eukaryotes has had their genomes sequenced (Gao et al. 2024; Miao et al. 2020). It was hoped that metagenomics may help solve this data shortage by taking advantage of any eukaryotic data that may be present in large metagenome libraries like Tara Oceans (Carradec et al. 2018), but some methodological issues and the greater complexity of eukaryotic genomes have stood in the way of fully achieving this (Eme and Tamarit 2024). First, abundant repeats, non-coding regions and large genome sizes in eukaryotes are challenging for current bioinformatic tools (Alexander et al. 2023; Eme and Tamarit 2024). Second, eukaryotes are usually less abundant than prokaryotes (Laforest-Lapointe and Arrieta 2018). Third, most existing bioinformatics tools are geared towards prokaryotic data (Alexander et al. 2023). For example, the reference databases for marker gene annotations (e.g., CheckM (Parks et al. 2015) and CheckM2 (Chklovski et al. 2023)) or taxonomy classifications (e.g., GTDB-tk (Chaumeil et al. 2020)) are specific to prokaryotes (Chaumeil et al. 2020; Chklovski et al. 2023; Kutuzova et al. 2024; Pan et al. 2023; Wang et al. 2024). Despite these hindrances, recent work has recovered hundreds of eukaryotic MAGs from shotgun metagenomes (Alexander et al. 2023; Delmont et al. 2022; Duncan et al. 2022; Patin and Goodwin 2022). The taxa for which genomes have been recovered, however, are often those with small genomes like prasinophytes (Alexander et al. 2023; Duncan et al. 2022), and any MAGs recovered from other eukaryotic taxa are largely incomplete (Krinos et al. 2024), hindering comparative genomic analyses (Duncan et al. 2022).

One potential path forward may be to focus on the organellar (plastid and mitochondrial) genomes of eukaryotes. Plastids, for example, are remnants of the ancestral cyanobacteria involved in the ancient endosymbiosis events that gave rise to the organelles (Sibbald and Archibald 2020). Through endosymbiotic gene transfer (EGT) and gene loss, the gene content of plastid genomes has become highly streamlined (Sibbald and Archibald 2020), making them significantly smaller than many prokaryote genomes. Usually, they have a higher copy number, and hence higher sequence coverage, when compared to nuclear genomes (Piganeau and Moreau 2007; Gallaher et al. 2018). They also commonly have a distinct nucleotide composition (Turmel et al. 2015), which may facilitate their assembly and binning into MAGs.

Organellar genes are frequent targets for phylogenetic analysis, with plastid genes such as *rbcL* and *tufA* frequently used in species delimitation and DNA barcoding (Borer et al. 2025; Sauvage et al. 2016), and plastid genome-scale trees offering great advances to our knowledge of algal evolution (Costa et al. 2016; Fang et al. 2017; Sun et al. 2016). Currently, ~2300 complete plastid genomes from ca. 1100 algal species are recorded in GenBank (2025), a small fraction of the over 50,000 algal species estimated to exist by 2024 (Guiry 2024). Metagenome-derived plastid MAGs (ptMAGs) could help mitigate the data sparsity and facilitate studying their evolution and genomic features. Early attempts successfully recovered plastid genomes

from Sargasso Sea metagenomes with high sequence coverage (Piganeau and Moreau 2007). Such motivation already led to bioinformatic developments, including *plastiC* (Cameron et al. 2023): the only workflow targeting metagenomic plastid sequences, by clustering contigs from metagenomes via MetaBAT2 (Kang et al. 2019) and keeping putative plastid bins (predicted by Tiara (Karlicki et al. 2022)) without prior filtering of assemblies. Another team proposed a GRM pipeline incorporating human-guided binning from *anvi'o* workflow (manually grouping contigs into bins based on k-mer and coverage (Eren et al. 2015)), which does not easily scale to larger metagenome datasets (Karlicki 2024). The results from these recent attempts offer valuable data on the feasibility of extracting ptMAGs from metagenomic data. However, handling large volumes of data in metagenomics requires scalable workflows and automation, and a plastid genome-specific database of marker genes to assess MAG quality (Parks et al. 2015) and guide binning (Hickl et al. 2022) is missing in existing approaches.

In this study, we present ChloroScan, an automated computational workflow that targets plastid genomes in metagenomes. It addresses the above challenges by automating targeted binning, taxonomic prediction, and quality control, guided by a manually curated database of plastid genes. ChloroScan generates information-rich summaries of the obtained ptMAGs to help interpret and use them in downstream applications. We benchmark ChloroScan against currently used tools using simulated and real metagenome data.

2 | Materials and Methods

2.1 | ChloroScan Workflow Overview

ChloroScan is a Snakemake-based (Mölder et al. 2021) workflow with the command line interface wrapped by *snk* (Wirth et al. 2024), to infer ptMAGs from metagenome contigs, with the utilities mostly written in Python and Unix bash. It is underpinned by a deep learning-based module to predict contigs of plastid origin and a manually curated database to guide metagenome binning and quality assessment (Figure 1).

One of the initial inputs of ChloroScan is the assembled contigs from large-scale metagenomic sequencing, which may contain contigs from bacteria, archaea, viruses, fungi, or phytoplanktonic algae. Hence, the first step in ChloroScan classifies contig categories using a deep learning contig classifier (Corgi (Turnbull 2025)) to denoise the assembly and offer a plastid-enriched assembly for downstream processes. For each contig, Corgi gives a probability of the target sequence being one of five RefSeq categories: eukaryotic nuclear, mitochondrial, plastid, bacterial, archaeal, and unknown otherwise. If the probability of one sequence type exceeds all others and reaches a given threshold, the sequence is assigned to that category. By default, ChloroScan runs Corgi with a probability threshold of ≥ 0.50 for the plastid category, and retaining only contigs ≥ 1000 bp, but these settings can be altered to optimise data mining performance in different data contexts.

The second step, metagenome binning, is a key step of the ChloroScan workflow. We employ *binny* (Hickl et al. 2022),

ChloroScan: recover plastid bins in one go

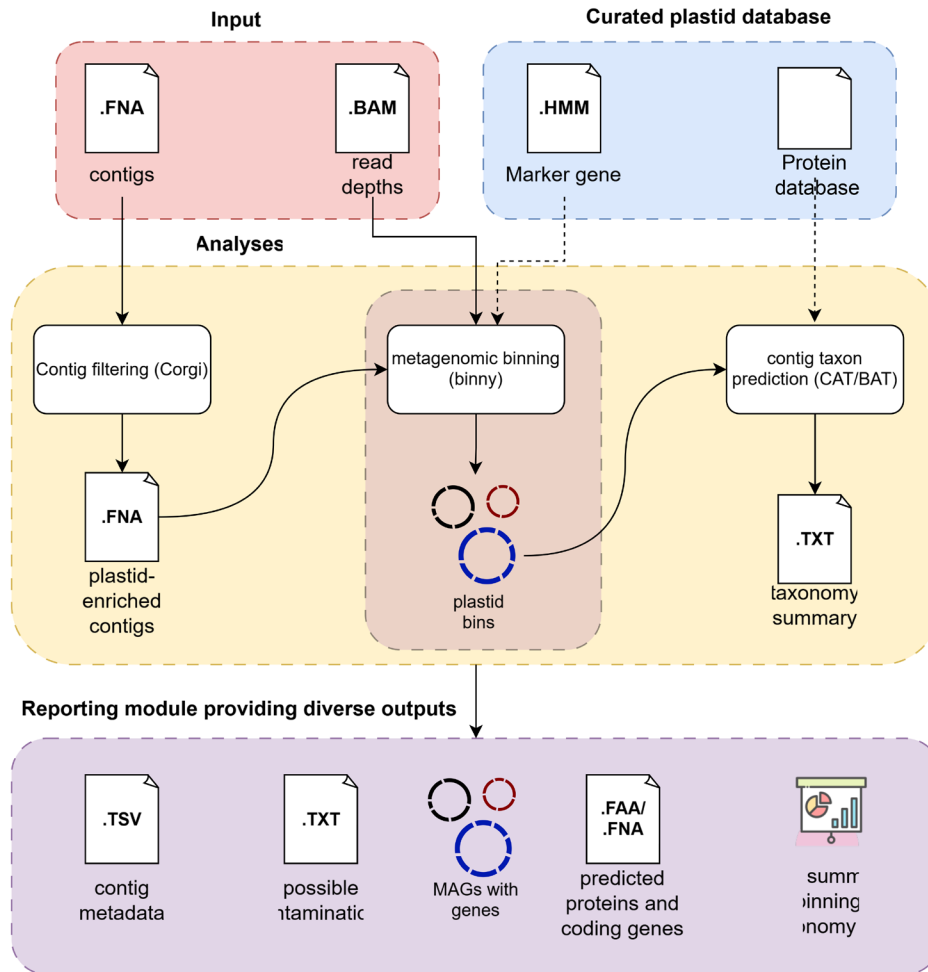


FIGURE 1 | ChloroScan's workflow structure. ChloroScan contains the following modules: Contig prediction by Corgi (Turnbull 2025), binning by binny (Hickl et al. 2022), taxonomic prediction by CAT/BAT (Von Meijenfeldt et al. 2019), and a summary module that generates user-friendly information to investigate the contig and bin data, including plots to investigate bin homogeneity, a table with contig metadata and the predicted genes and proteins from MAGs. It takes assemblies in FASTA format and sorted BAM files as raw inputs. It is configured by a custom marker gene database in CheckM's format (Parks et al. 2015) and a protein database that works for binning and taxonomic predictions, respectively. Normal arrows represent passing files as inputs for steps and dotted lines represent passing configuration files or directories to the steps. The binning module (enclosed by a shaded box) is further engineered to target plastid genomes rather than prokaryotes.

which uses k-mer frequencies and sequence abundance in clustering contigs and uses marker genes to guide the clustering of contigs into bins. Binny's original database to guide clustering focuses on prokaryotes (CheckM (Parks et al. 2015)). Hence, we designed a custom database for plastid marker genes following its rationale (Parks et al. 2015; see [Supporting Information](#)), transforming it to recover plastid bins. We altered binny to not run read depth calculations, but rather do this more efficiently within the core ChloroScan workflow. For the recovery of single-contig MAGs (scMAGs), we set default thresholds of 30 for marker gene count and 85% for marker completeness.

Following binning, taxonomic classification adds an extra layer of information for users to interpret their data by showing contig-level taxonomy identification. ChloroScan uses CAT/BAT (Von Meijenfeldt et al. 2019) with default settings to predict the taxa of contigs based on the ORFs it predicts and searches the ORFs against a protein database to infer the likely taxonomic origin of contigs. For this step, we designed a custom protein reference database that includes the Uniref90 (Suzek et al. 2007) database to provide taxonomic breadth along with an extensive custom database of plastid-encoded protein sequences of algae to enhance the resolution of taxonomic identifications. This dataset replaces the original CAT/BAT-prepared nonredundant protein database

(nr), which requires more download and storage capacity, run-time, and memory usage.

After these computations, ChloroScan uses a range of custom Python scripts to deliver diverse user-friendly outputs and summaries that enable users to make sense of their data and use the results for a range of downstream applications. The first step in this section is to detect contamination in bins. Due to the lack of published refinement tools for ptMAGs, we adopted a conservative contamination detection based on contig-level taxonomy prediction by CAT/BAT to highlight bins containing contigs with ambiguous or non-eukaryotic taxon predictions and without ORFs matching our plastid marker gene database. We also provide community-level taxonomic information of the Corgi-filtered metagenome with Kronatools (Ondov et al. 2011). ChloroScan produces a range of graphs, including a scatterplot of GC contents versus Log10 transformed average coverage showing the homogeneity of bins, pie charts showing the taxonomic assignments of the contigs in each bin, and a violin plot demonstrating the contig coverage distribution for each bin. A spreadsheet reports detailed contig statistics, including their length, marker genes, and taxon prediction.

Finally, ChloroScan predicts the coding nucleotide sequences and proteins of bins. Due to the potential occurrence of gene fragments in complex metagenome assemblies, we chose FragGeneScanRs (Van der Jeugt et al. 2022). The coding sequences and proteins of each bin are provided in FASTA format, which can directly feed into downstream comparative and/or phylogenomic analyses. The location of genes on the contigs is also given as a GFF3 file.

2.2 | Benchmarking ChloroScan With Synthetic Data

ChloroScan was benchmarked using simulated marine metagenomic datasets. We designed simulated samples containing prokaryotic, eukaryotic nuclear, eukaryotic mitochondrial, and plastid genomes. We generated two simulated samples with different sets of relative abundances of each genome using CAMISIM (Fritz et al. 2019; CAMI-challenge, n.d.), with an N50 of 17,762bp for sample 1 and 15,843bp for sample 2; the simulation method is described in [Supporting Information](#). These samples contain 12 plastid genomes: 11 canonical plastid genomes and one dinoflagellate plastid genome characterised by minicircles. We generated read depth profiles by aligning the simulated reads to the metagenome assemblies using minimap2 version 2.26 (Li 2018). As we failed to run *plastiC* ([Supporting Information](#)), we used its underlying binner *MetaBAT2* (Kang et al. 2019) for benchmarking comparison. When running ChloroScan, we set the contig length cutoff to 1500bp, to match *MetaBAT2*'s minimum contig length cutoff. Other settings remained default, and we set the default minimum completeness of bins to 50% to recover as many bins as possible despite their sizes. Binning results from these tools and ground truth were analysed using AMBER (Meyer et al. 2018), including calculations of the F1 score, a summary performance metric combining precision, recall, bin purity and completeness and accuracy ([Supporting Information](#)). Sequence coverage is known to impact genome recovery. To help explore the relationship between

ChloroScan's recovery outcome and source genomes' coverage, we generated mean read depth for 11 canonical source genomes via mapping simulated reads to these genomes using minimap2 and running depth calculation steps from binny's workflow. To assist in checking the relative abundance of each genome, we computed them using CoverM's "genome" pipeline (Aroney et al. 2025). The corresponding data are in File [S1](#).

2.3 | Application to Real Metagenomes

We then applied ChloroScan to four marine metagenomes from the Tara Oceans expedition datasets, using the three size fractions under the category "protist": SAMEA2189670 (0.8–5µm), SAMEA2732360 (180–2000µm), SAMEA2657032 and SAMEA2732613 (20–180µm). Most samples were collected from surface sea waters, except SAMEA2732613 at 40m depth. We downloaded their MEGAHIT assemblies from SPIRE (Schmidt et al. 2024), consisting of ~24M contigs. We then mapped the samples' raw reads using minimap2 (version 2.26) in short-read mode. Due to a lack of prior knowledge of these samples, to assess those medium and high-quality ptMAGs with more confidence of their homogeneity, we chose a conservative set of parameters while running ChloroScan, by setting minimum completeness and purity cutoff to 70% and 90% (The Genome Standards Consortium et al. 2017). Other settings remained as default. Later, to inspect the resulting bins' taxonomy, we retrieved their *rbcL* genes via Orthofisher (Steenwyk and Rokas 2021; see [Supporting Information](#)) and blasted them against a nonredundant protein database.

3 | Results

3.1 | ChloroScan Recovers High-Quality ptMAGs From Synthetic Metagenomes

To compare ChloroScan's performance against that of similar software, we benchmarked it alongside *MetaBAT2*, the binner used by *plastiC* (Cameron et al. 2023). Both solutions produced ptMAGs (Figure 2), but we found that those produced with ChloroScan had higher overall quality and purity, with the average F1 score at sequence level and accuracy for the two simulated samples 31.3% and 23.7% higher than *MetaBAT2*, respectively (Tables [S1](#) and [S2](#)). In total, ChloroScan recovered 8 and 9 bins from these two synthetic metagenomes, with 6 and 8 of these high-quality according to MIMAG (The Genome Standards Consortium et al. 2017). Five and six were near-complete single-contig MAGs. *MetaBAT2* recovered 6 and 7 bins, of which only 3 were considered high quality (The Genome Standards Consortium et al. 2017) in both samples (Figure [2c,d](#)). For sample 1, the remaining plastids were also retrieved as fragmented genome sequences by ChloroScan, with completeness ranging from 2.2% to 94.5% (see File [S1](#)), resulting in an overall accuracy of 0.848 and an F1 score of 0.92. *MetaBAT2*'s accuracy was 0.753 and F1 score of 0.573 for this sample (Table [S1](#)). For sample 2, the plastid genomic sequences are less fragmented than in sample 1: the completeness ranges from 73.2% to 99.9% (see File [S1](#)). ChloroScan's accuracy (0.878), F1 (0.735) and average purity per bp (0.982) compared favourably to *MetaBAT2*'s corresponding values (Table [S2](#)).

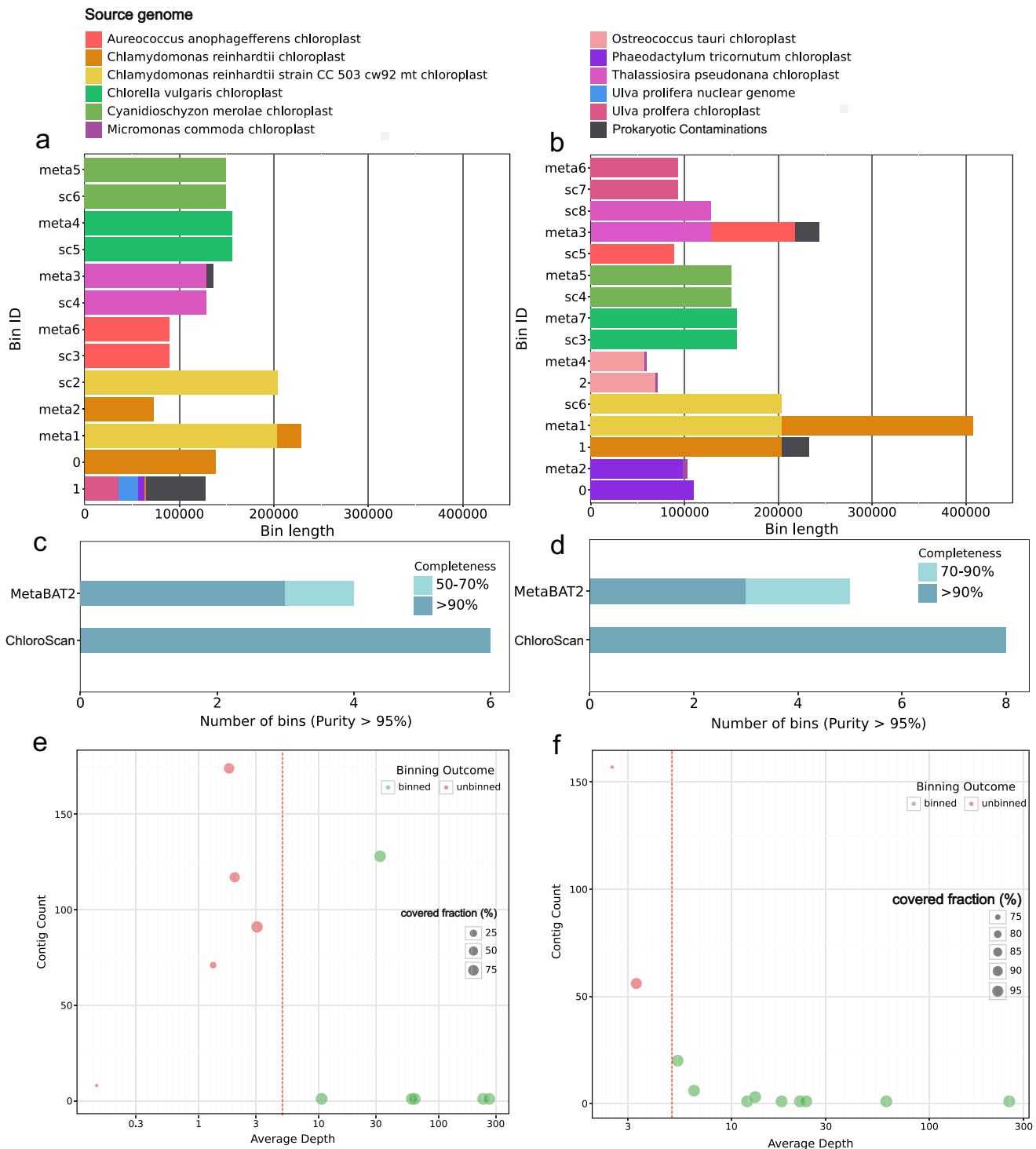


FIGURE 2 | Plots of ChloroScan results show its effectiveness compared to MetaBAT2. The bar charts compare bins from ChloroScan to bins from MetaBAT2 in (a) single sample metagenome 1 and (b) single sample metagenome 2, in terms of their homogeneity by showing how many taxa are included in one bin. ChloroScan bins are labelled as digits and MetaBAT2 bins are prepended with “meta”. Single contig MAGs recovered by ChloroScan are labelled as “sc”. Prokaryotic contigs regardless of their taxa are labelled dark grey. The stacked bar charts on the right side demonstrate the count of ptMAGs in different quality classes from each tool in the (c) sample 1 and (d) sample 2. The relationship between the average read depth calculated by binny’s workflow, the level of fragmentations of each source genomes corresponding and their recovery status is plotted for sample 1 (e) and sample 2 (f). The red dashed vertical lines represent an average depth of 5x.

For sample 1, both tools had a similar adjusted random index (ARI): a summary metric to measure the performance of clustering, but ChloroScan recovered bins from a higher percentage of binned nucleotides than MetaBAT2 (Figure S1a).

For sample 2, ChloroScan’s ARI substantially exceeded MetaBAT2’s (Figure S1b). The detailed bin taxonomic compositions show that MetaBAT2 joined the plastid genomes of two *Chlamydomonas reinhardtii* strains (meta1) and of

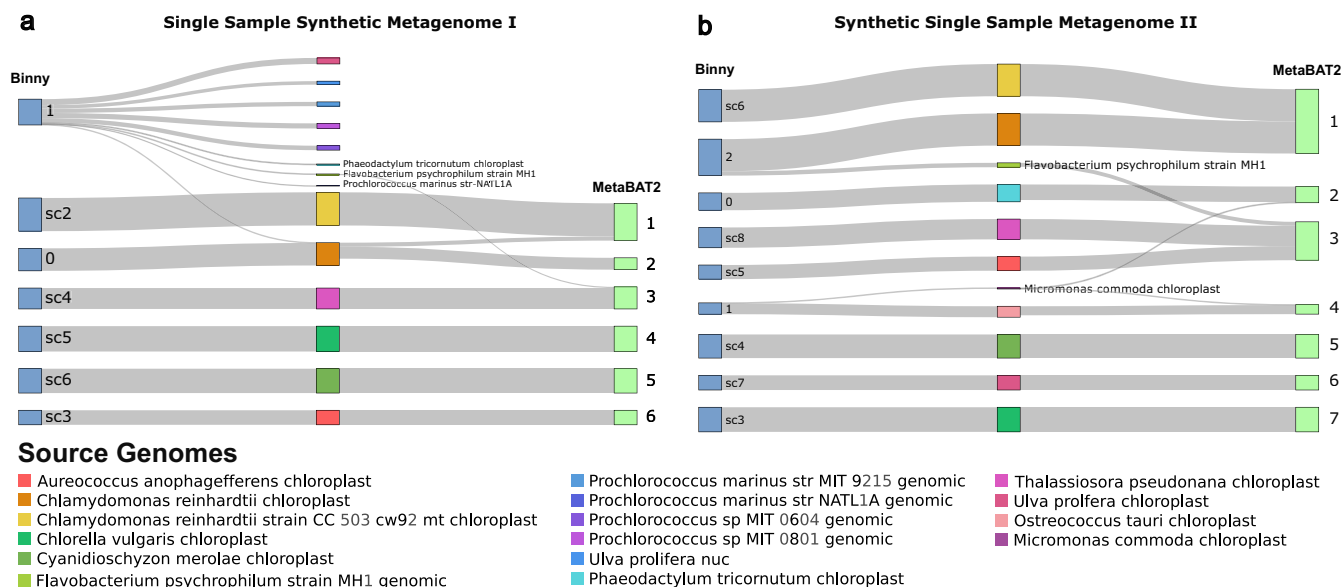


FIGURE 3 | Mapping information from each bin to source genomes in the (a) synthetic single-sample metagenome 1 and (b) 2 based on the contig mapping information generated from CAMISIM. Grey bar widths refer to the percentage of source genome length taken by contigs. Here one contig has only one source genome mapped. Source genomes with too short contigs (colours invisible in the Figure) in the sample have their names labelled near the bar. Binny-recovered single-contig MAGs are prepended with “sc”.

two distinct heterokont species: *Thalassiosira pseudonana* and *Aureococcus anophagefferens* (meta3; Figures 2c,d and 3b). Except those single-contig MAGs (with prefix “sc” in Figure 2a,b), the multi-contig ptMAGs from sample 1 show wide variation of GC contents, with the ChloroScan-recovered bin 0 having coverage around 30× and the bin 1 having coverage around 3× with more variations in GC contents. In sample 2, the clusters appear coherently in the scatterplot, with nearly uniform read depths within MAGs and less GC content variation compared to synthetic metagenomic sample 1 (Figure S2a,b). Neither sample recovered dinoflagellate bins consisting of minicircular chromosomes, and the Corgi-filtered assemblies from two samples did not contain their minicircular chromosomes (Figure S3).

To better understand the average depth threshold above which MAGs can be recovered, we produced the average read depth of source genomes and generated scatterplots to visualise relationships between average depth and fragmentation (quantified as the count of contigs) of the 11 canonical source genomes (Figure 2e,f). We found that in sample 1, the five unbinned genomes have sequence mean coverage of less than 5×. The assembly of *Micromonas commoda* consists of only 8 contigs and only covers a small fraction of the complete genome, so it is the assembly that failed to recover its genome. The other four unbinned genomes’ contigs cover a larger fraction of the genome, but with a high level of fragmentation. One exception is *Chlamydomonas reinhardtii* (bin 0, see Figure 2a). Despite the high level of fragmentation, it has coverage around 30× and was still recovered with desired completeness and purity. In contrast to sample 1, the quality of binning from sample 2 is significantly better than sample 1, with overall higher coverages across all sampled genomes. Only two source genomes (*Isochrysis galbana* and *M. commoda*) were not recovered, both having a fragmented assembly and a coverage less than 5×.

Importantly, ChloroScan bins have higher average completeness than MetaBAT2 bins (Figure S1c,d), and the majority of ChloroScan bins were nearly complete and without any contamination (Figure 2a,b). One important exception is the highly multimeric bin (bin 1) produced by ChloroScan for sample 1 (Figures 2a and 3a), which led to the average purity score for ChloroScan bins being lower than that for MetaBAT2 (Figure S1c). This multimeric bin is made up of short contigs from different plastid, nuclear, and prokaryote genomes with overall low coverage around 3× and diverse GC contents (Figure S2a) and was clearly indicated as a taxonomic mosaic based on the contig identifications provided by ChloroScan.

3.2 | 16 ptMAGs Recovered From Four Ocean Metagenome Samples

ChloroScan’s ability to recover plastid genomes from real datasets was clearly illustrated by its application to four Tara Oceans marine metagenomes. Out of ~24 million contigs, ChloroScan classified 16,556 putative plastid contigs and the binning module inferred 16 ptMAGs with completeness > 70% and purity > 90% (medium to high quality according to MIMAG (The Genome Standards Consortium et al. 2017)). The scatterplots of two samples (SAMEA2189670 and SAMEA2732360) demonstrate that most bins have homogeneous coverage and slight variation in GC content (Figures 4a and S4).

To demonstrate ChloroScan’s ability to recover novel MAGs, we looked at taxonomic predictions for eight bins from the sample SAMEA2732360 (collected from Antarctic coast) and the two bins from SAMEA2189670 (collected from southern Mediterranean), based on the CAT taxonomic predictions implemented in ChloroScan (Figure 4b) and BLASTN search

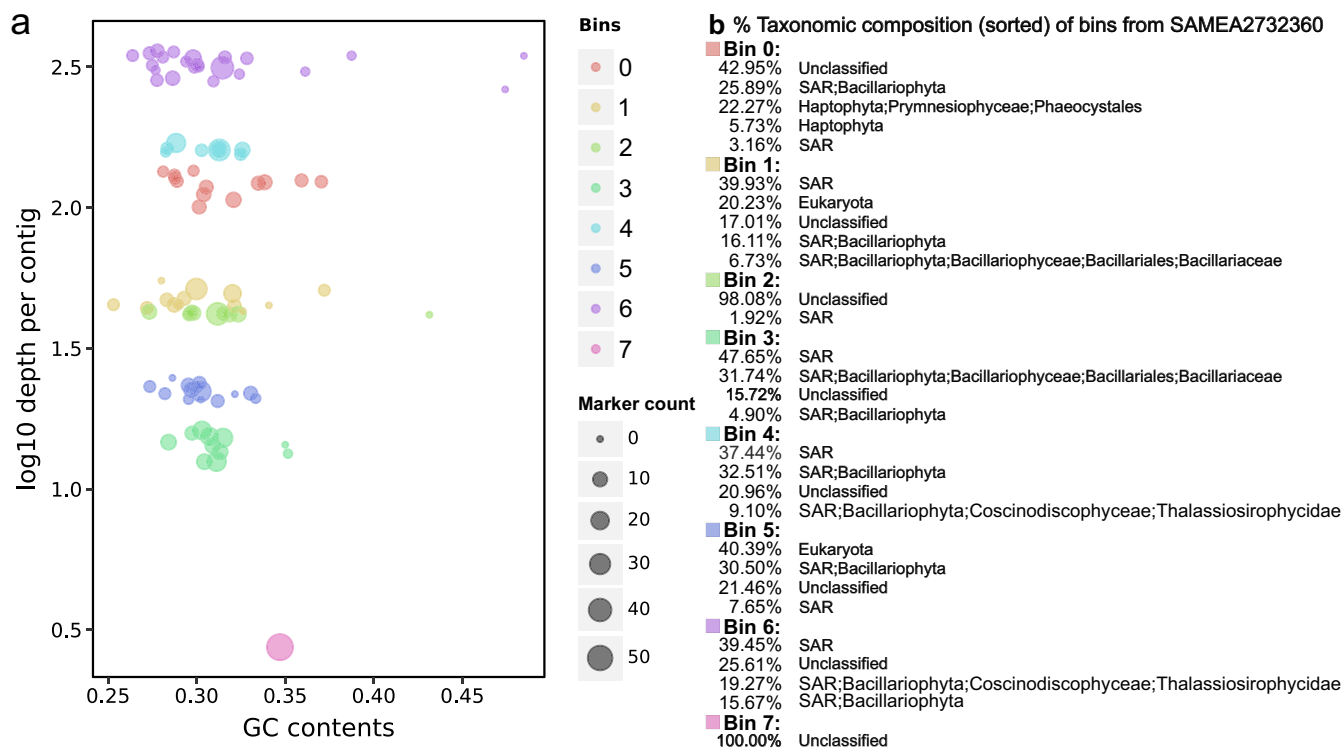


FIGURE 4 | Metagenome-assembled genomes from real marine metagenomes. (a) The GC x log10 average read depth plots of the sample SAMEA2732360. Marker gene count per contig is scaled by the dot size. **(b)** Contig-level taxonomy composition of six bins from the sample SAMEA2732360 inferred by CAT. The sorted percentages of MAG length taken by each taxon are listed on the left side, and the detailed taxon lineages (adapted from NCBI taxonomy) for eukaryotic contigs are on the right side. The “Unclassified” refers to all categories without an exact taxon name (i.e., Root, unclassified, cellular organisms). The plot for SAMEA2189670 is in [Supporting Information](#) (Figure S4).

of the marker gene *rbcL* against core nonredundant nucleotide database (nt). Overall, diatom species are prevalent in SAMEA2732360. Bins 1–6 show largely nested taxonomic composition predicted by CAT, with contigs either unclassified, belonging to the SAR supergroup: the group that diatoms belong to, or with more specific identity. But there are also exceptions. Bin 0 contains some putative haptophyte contigs in addition to putative SAR contigs, potentially representing slight chimerism. Bin 7 is not classified based on our protein database containing all protistan plastid proteins and Uniref90 proteins. Our BLASTN results using the *rbcL* gene recovered from each of these bins provided more fine-grained taxonomic identifications. For Bin 0, we got strong hits to several diatom species mainly from Coscinodiscophyceae (Supporting Information S1). For Bin 1, the highest hit (94.84%) is with *Fragilariopsis kerguelensis*—a pennate diatom native to Southern Ocean with one of the highest abundances in the sediments (Warnock et al. 2015; see Supporting Information S2). Considering the <95% similarity, it seems more likely our bin represents another species within the genus. Bin 2 resembles *Chaetoceros danicus*, with 98.29% *rbcL* similarity (Supporting Information S3). *Pseudo-nitzschia turgiduloides*—a pennate diatom, matches the *rbcL* sequence of bin 3 at 99.52% identity (Supporting Information S4). Bin 4 shows the best hit (95.11%) with the diatom *Eucampia zodiacus*, a species known to cause harmful algal blooms and widely distributed in non-polar waters (Zhang et al. 2021; Supporting Information S5). Our work likely identified a close relative of it that exists in polar waters. Bin 5's identity is narrowed to *Chaetoceros*,

most similar to *Chaetoceros gelidus* with 99.93% similarity (Supporting Information S6). Bin 6's *rbcL* had *Corethron hystrix* as the top BLASTN hit (93.55%), indicating relatedness but not that exact species (Supporting Information S7). It has the highest abundance among all bins (Figure 4a). Finally, bin 7's *rbcL* is identical to a eudicot (muskmelon), which is likely a sample/lab contaminant (Supporting Information S8). In sample SAMEA2189670, the bin 0 contigs have a nested SAR origin predicted by CAT. The *rbcL* blast suggests it as *Pseudo-nitzschia cuspidata* from the family Bacillariaceae (class Bacillariophyceae), with similarity ~97% and query cover 96%. Other taxa in top hits also belong to the same class, giving a putative Bacillariaceae origin (Supporting Information S9). Finally, the single-contig MAG (with length 81,223 bp) bin 1 from SAMEA2189670 has predicted proteins that resemble those of ochrophytes. The BLAST search does not give close hits, rather 85%–87% similarity with several distantly related freshwater chrysophytes (Supporting Information S10), suggesting that this bin is likely a new deep-branching lineage within Ochrophyta.

The Krona plots (Figure S5) demonstrate that Stramenopiles and pico-sized green algae (e.g., *Chloropicaceae*) dominate the samples, with Bacillariophyta (diatoms) and Ochrophytes dominating. Haptophytes are also found in SAMEA2189670 with the smallest size fraction. Some rare taxa such as Cryptophytes, Discoba and the unicellular red algae family *Galdieriaceae* also exist in the smaller size fraction. In addition, Corgi included a fraction of non-photosynthetic metazoan sequences into the

putative plastid contig assemblies for all samples. Overall, most contigs with taxon assigned are restricted to higher ranks (e.g., phylum, class and order) and a considerable fraction of contigs were not assigned to any taxon.

4 | Discussion

We developed ChloroScan, a metagenomic binning workflow targeting plastid genomes, and showed its performance using synthetic and real metagenomes. ChloroScan leverages an existing binning framework designed for prokaryotes (Hickl et al. 2022), but we enhanced its performance for plastid genome binning with a manually designed plastid-encoded marker gene database and settings fine-tuned to retrieve plastid genomes. The utility of ChloroScan to recover plastid genomes from real metagenomes was illustrated by recovering 16 ptMAGs from just four marine metagenome samples (Figures 4 and S4).

Our work offers several innovations in the field. Previous works often used manual binning to sort plastid contigs (Jamy et al. 2025), while our automated workflow offers potential for screening much larger datasets. An important innovation in ChloroScan is that it uses plastid marker genes to improve ptMAG recovery. This approach was known to work well in recovering prokaryotic genomes (Hickl et al. 2022), and our results extend this to plastid genomes (Figures 2a and 3b). The current version of this database is built from 458 reference plastid genomes. With more sequenced reference genomes added to analyses with appropriate filtering for those non-canonical plastids or highly reduced plastids, we believe our database's resolution will be improved through the iterative updates to this database. Increased ptMAG sampling is also promising to improve this database's resolution, as shown by GTDB, with more MAGs than isolates' genomes in Archaea (Parks et al. 2022), which then contributed to the model training of CheckM2 in providing more accurate MAG quality estimates (Chklovski et al. 2023). But the MAG qualities should be ensured through careful selections. ChloroScan also outperformed MetaBAT2 in accurately separating two strains of *C. reinhardtii* from the synthetic metagenome, suggesting that it can separate strain-level plastid genomes in complex metagenomes, with enough sequence coverage. To sort out low-quality bins, ChloroScan's visual reports allow the user to easily identify bin quality and possible chimeras as a primary sanity check. An additional novelty we present is incorporating organelle genomes into metagenomic simulations. Current simulated metagenomes are often prokaryote-oriented, with only a few examples containing eukaryote genomes (Marcelino et al. 2020). Our approach went well beyond that, with extensive sampling of eukaryote mitochondrial, plastid, and nuclear genomes.

Our simulations show that low coverage is a logical limitation that can severely impact the length of contigs, resulting in them being filtered out prior to binning, and missing the chance of recovering those rarer plastids (Figure 2e,f). The boundary appears to be at ca. 5× average coverage, with source genomes sequenced at lower coverage being more fragmented and less likely to be recovered with high purity, as exemplified by *Micromonas commoda* (Figure 2e,f). Meanwhile, we found that in sample 1 four highly fragmented genomes were not binned despite being

assembled at higher coverage and better completeness than *M. commoda*. Likely due to the contig length cutoff of ChloroScan used in this study (1000 bp), few contigs from these source genomes entered the binning stage, thus Binny failed to include them in binning despite their contigs covering nearly the whole source genomes. Binny was shown to outperform MetaBAT2 in recovering fragmented prokaryotic MAGs (Hickl et al. 2022). However, plastid genomes are shorter, and the presence of highly fragmented, low-coverage genomes would produce numerous contigs with biased nucleotide frequencies, making downstream binning more vulnerable to producing chimeras (Figure 2a).

The taxonomic predictions of ptMAGs recovered from four real marine metagenomes (Figures 4 and S4) are consistent with the knowledge from 16S plastid rRNA sequencing analysis that Chlorophyta, Haptophytes, and diatoms are some of the most abundant phytoplankton groups in the Tara Oceans samples (De Vargas et al. 2015; Pierella Karlusich et al. 2023; Penot et al. 2022). Many of these are small and they appear to have substantial biodiversity that is yet to be discovered (De Vargas et al. 2015). For example, prasinophytes, which we found in several samples (Figure S5), are dominant green algal taxa in surface oceans, with prasinophyte clade VII particularly taking high abundance in metabarcoding analysis of Tara Oceans data (Lopes Dos Santos et al. 2017).

Our taxonomic identifications using CAT/BAT and BLASTN searches of the *rbcL* gene showed: (1) clear evidence for the recovery of plastid genome bins of common ocean phytoplankton, and (2) plastid genomes from novel lineages. At genome-level, bins recovered with taxonomic identification from the polar sample SAMEA2732360 represent the commonly found species from marine microbiomes: diatoms and other ochrophytes, both from the SAR supergroup known to be dominant in Tara Oceans samples (Pierella Karlusich et al. 2023). Among recovered ptMAGs, we recovered a polar-representative MAG: bin 1 that resembles *Fragilariopsis kerguelensis*, which presents high abundance in Antarctic waters (Warnock et al. 2015). We also found ptMAGs of cosmopolitan taxa (e.g., *Pseudo-nitzschia* and *Chaetoceros*). Bin 5 from SAMEA2732360, has high *rbcL* similarity (>95%) to *Eucampia zodiacus* that is found worldwide except in polar waters. Hence, we substantially expand the knowledge regarding this species while offering candidates for future genomic comparisons. Additionally, we recovered an ochrophyte ptMAG that may come from a novel lineage. Bin 1 from SAMEA2189670 shows extremely high contiguity and has only distant matches (ca. 85% identity) among its top ten BLASTN hits of the *rbcL* gene, with hits of similar identity to several chrysophyte genera (Supporting Information S10). Seeing that chrysophytes are a nearly exclusively freshwater lineage (Debroas et al. 2017), we consider this lineage to likely be a yet-to-be-discovered marine ochrophyte lineage (see also Terpis et al. 2025). We consider the top BLASTN hit of the *rbcL* sequence (annotated as “uncultured bacterium” in NCBI) to be a misidentified algal plastid sequence containing *rbcL*.

Our results also show that ChloroScan did not recover dinoflagellate plastid bins (Figure S3) despite having high overall relative abundance (File S1). Dinoflagellates have an unusual plastid structure, consisting of several minicircles with each containing a single gene (Howe et al. 2008). In the simulated

datasets, the contigs corresponding to these minicircles were excluded from the filtered assembly by Corgi despite having high coverage and contiguity. Corgi was trained on RefSeq genomic sequences to predict contig categories, and of the handful of dinoflagellate mini-circle plastid chromosomes currently known; none were found in RefSeq's plastid genome category (O'Leary et al. 2016). Hence, Corgi assigns dinoflagellate plastid sequences with relatively low probability of being in the "plastid" category, but this situation could arguably be rectified in future by refining the curation of the training data for Corgi. Previous 16S rRNA sequencing results (Pierella Karlusich et al. 2023; Penot et al. 2022) show that dinoflagellate plastid 16S rRNA has very low relative abundance in real metagenomes. This may indicate that their plastid genomes may also be sequenced at low coverage, lowering the probabilities of successful recovery of their ptMAGs from real metagenomes, but it should also be considered that dinoflagellate 16S rRNA are highly divergent (Pierella Karlusich et al. 2023) which may hinder their detection in the survey. Other than dinoflagellates, some other protist lineages (e.g., apicomplexans) and siphonocladous algae, also have highly reduced plastid genomes with non-canonical structures (Del Cortona et al. 2017; Sato 2011), which our current workflow and curated marker gene database are not currently tuned for. But they offer opportunities for future developments.

The Corgi contig classification module inevitably retains some contamination in the filtered assembly (Figure S3), causing potential downstream prokaryotic contamination for the binning module to contend with, as shown in our benchmarking data. The taxonomic reports from CAT/BAT (Von Meijenfeldt et al. 2019) and visualisations offer a chance for users to inspect and remove such contaminant contigs. Our workflow reports contaminant contigs without marker genes or non-eukaryotic predicted identity, which can help users refine their ptMAGs.

Plastid genomes are mainly present in photosynthetic protists (some have reduced or lost plastid genomes like apicomplexans); hence this approach is bound to miss heterotrophic microeukaryotes. However, analogous approaches using mitochondrial genomes might be proposed. Insect studies (Crampton-Platt et al. 2016) suggest that this is feasible, and a survey from 2019 revealed substantial mitochondrial genome diversity of flagellated protists from Pacific Ocean samples (Wideman et al. 2019), making it an appropriate target to sample undiscovered heterotrophic protists. ChloroScan is flexible to adapt to mitochondrial MAG recovery, as Corgi also predicts mitochondrial contigs, but a curated database of mitochondrial marker genes from each lineage would need to be designed and implemented to achieve this.

While the metagenomic bioinformatics development for eukaryotic plastids has just begun, relevant package developments to handle assembled reference plastid genomes are also proceeding. CPStools (Huang et al. 2024), which takes the plastid genomes in GenBank or FASTA format, assembles 10 functionalities into one workflow, including predicting CDS, annotating genomic structures, converting files into different formats, and calculating relevant statistics from newly sequenced plastid genomes. Unlike CPStools, ChloroScan focuses on noisier unassembled contigs, which are less informative about plastid genome structure and potentially have more fragmented genes to predict. But

both tools could provide a standardised pipeline for handling data from either reference genomes or MAGs, and combining both tools will ease plastid-related bioinformatic analyses, providing more high-quality data and yielding more valid results.

In summary, the ChloroScan workflow facilitates automated binning of ptMAGs from filtered contigs, with the use of a marker gene database to improve binning sensitivity and accuracy. It accurately predicts plastid contigs and recovers ptMAGs with higher completeness and purity from synthetic and real metagenomes. As the availability of reference plastid genomes continues to grow, the marker gene database and Corgi training data will capture a broader range of taxa and enhance the precision of binning and taxonomic identifications. As more studies attempt to recover ptMAGs from metagenomes (Jamy et al. 2025), the resulting MAGs can serve as valuable resources to address a variety of downstream biological questions and further bioinformatic advances. With progress made in MAG recovery and reference genomes handling, we believe that the volume of high-quality eukaryote plastid genome data could burst in the upcoming years and finally provide a more holistic map of its evolution and contribute to other areas.

Author Contributions

Yuhao Tong contributed to workflow construction and adjustment, metagenome data simulation, GitHub repository maintenance, and data analysis. Robert Turnbull contributed to technical support while doing bioinformatic developments and Corgi's training and maintenance. Heroen Verbruggen and Vanessa Rossetto Marcelino contributed to workflow conceptualisation of ChloroScan. All authors contributed to manuscript writing and revision.

Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services. H.V. is supported by a fellowship from the Fundação para a Ciência e a Tecnologia (CEECIND:2023.06155). V.R.M. was supported by the Australian Research Council (DE220100965) and the Spanish Ministry of Science and Innovation (RYC2023-042907-I). Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australasian University Librarians.

Funding

This work was supported by The University of Melbourne's Research Computing Services, Fundação para a Ciência e a Tecnologia (CEECIND:2023.06155), Australian Research Council (DE220100965), Ministerio de Ciencia e Innovación (RYC2023-042907-I).

Disclosure

Benefit Sharing: The study has no benefits to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

ChloroScan is available in GitHub: https://github.com/Andyargueasae/chloroscan/tree/release_v0.1.7. Beginners' guide and utilities of ChloroScan can be found at: <https://andyargueasae.github.io/chloroscan/index.html>. The CAMISIM used in this study can be found at

<https://github.com/CAMI-challenge/CAMISIM/tree/dev>. The CAT database used in the current version of ChloroScan is deposited here in figshare: <https://doi.org/10.26188/27990278>.

The assemblies and bam files of two synthetic samples used in this study are available in figshare <https://doi.org/10.26188/28748540>. Real metagenomes' raw reads (SAMEA2189670, SAMEA2732360, SAMEA2657032 and SAMEA2732613) are available in ENA (<https://www.ebi.ac.uk/ena/browser/home>). Their assemblies assembled by spire workflow can be downloaded in the SPIRE database (<https://spire.embl.de/>), under the study name "TARA_Oceans_protists_metaG". The data we used from these four metagenomes (assembly and bam files) are dumped in figshare: <https://doi.org/10.26188/31573732>.

The OrthoFinder results on picked genomes, the original and intermediate data to generate the marker gene database, the outputs from ChloroScan for all benchmarking metagenomes (sample 1 and 2) and four real metagenomes, the *rbcL* hits from bins from SAMEA2732360 and SAMEA2189670, and the nucleotide coding sequences of these bins are deposited in figshare: <https://doi.org/10.26188/28722788>. Detailed instructions and codes to construct the plastid marker gene database for binny and all intermediary files are saved in the GitHub repository: https://github.com/Andyargueasae/ChloroScan_reproducibility.git.

When downloading files from the figshare articles, due to aws bot detection that blocked wget and curl from access, we recommend using pyfigshare to download these files. Relevant tutorials can be found at the README.md page of ChloroScan: https://github.com/Andyargueasae/chloroscan/tree/release_v0.1.7.

References

Alexander, H., S. K. Hu, A. I. Krinos, et al. 2023. "Eukaryotic Genomes From a Global Metagenomic Data Set Illuminate Trophic Modes and Biogeography of Ocean Plankton." *MBio* 14, no. 6: e01676-23. <https://doi.org/10.1128/mbio.01676-23>.

Aroney, S. T. N., R. J. P. Newell, J. N. Nissen, A. P. Camargo, G. W. Tyson, and B. J. Woodcroft. 2025. "CoverM: Read Alignment Statistics for Metagenomics." *Bioinformatics* 41, no. 4: btaf147. <https://doi.org/10.1093/bioinformatics/btaf147>.

Borer, G., C. Monteiro, F. P. Lima, and F. M. S. Martins. 2025. "Performance of DNA Metabarcoding vs. Morphological Methods for Assessing Intertidal Turf and Foliose Algae Diversity." *Molecular Ecology Resources* 25, no. 7: e14115. <https://doi.org/10.1111/1755-0998.14115>.

Brown, C. T., L. A. Hug, B. C. Thomas, et al. 2015. "Unusual Biology Across a Group Comprising More Than 15% of Domain Bacteria." *Nature* 523, no. 7559: 208–211. <https://doi.org/10.1038/nature14486>.

Cameron, E. S., M. L. Blaxter, and R. D. Finn. 2023. "plastiC: A Pipeline for Recovery and Characterization of Plastid Genomes From Metagenomic Datasets." *Wellcome Open Research* 8: 475. <https://doi.org/10.12688/wellcomeopenres.19589.1>.

CAMI-challenge. n.d. "CAMISIM [GitHub Repository]." *GitHub*. <https://github.com/CAMI-challenge/CAMISIM/tree/dev>.

Carradec, Q., E. Pelletier, C. Da Silva, et al. 2018. "A Global Ocean Atlas of Eukaryotic Genes." *Nature Communications* 9, no. 1: 373. <https://doi.org/10.1038/s41467-017-02342-1>.

Chabé, M., A. Lokmer, and L. Ségurel. 2017. "Gut Protozoa: Friends or Foes of the Human Gut Microbiota?" *Trends in Parasitology* 33, no. 12: 925–934. <https://doi.org/10.1016/j.pt.2017.08.005>.

Chaumeil, P.-A., A. J. Mussig, P. Hugenholtz, and D. H. Parks. 2020. "GTDB-Tk: A Toolkit to Classify Genomes With the Genome Taxonomy Database." *Bioinformatics* 36, no. 6: 1925–1927. <https://doi.org/10.1093/bioinformatics/btz848>.

Chklovski, A., D. H. Parks, B. J. Woodcroft, and G. W. Tyson. 2023. "CheckM2: A Rapid, Scalable and Accurate Tool for Assessing

Microbial Genome Quality Using Machine Learning." *Nature Methods* 20, no. 8: 1203–1212. <https://doi.org/10.1038/s41592-023-01940-w>.

Costa, J. F., S.-M. Lin, E. C. Macaya, C. Fernández-García, and H. Verbruggen. 2016. "Chloroplast Genomes as a Tool to Resolve Red Algal Phylogenies: A Case Study in the Nemaliales." *BMC Evolutionary Biology* 16, no. 1: 205. <https://doi.org/10.1186/s12862-016-0772-3>.

Crampton-Platt, A., D. W. Yu, X. Zhou, and A. P. Vogler. 2016. "Mitochondrial Metagenomics: Letting the Genes Out of the Bottle." *GigaScience* 5, no. 1: 15. <https://doi.org/10.1186/s13742-016-0120-y>.

Dai, R., J. Zhang, F. Liu, et al. 2025. "Crop Root Bacterial and Viral Genomes Reveal Unexplored Species and Microbiome Patterns." *Cell* 188, no. 9: 2521–2539.e22. <https://doi.org/10.1016/j.cell.2025.02.013>.

De Vargas, C., S. Audic, N. Henry, et al. 2015. "Eukaryotic Plankton Diversity in the Sunlit Ocean." *Science* 348, no. 6237: 1261605. <https://doi.org/10.1126/science.1261605>.

Debroas, D., I. Domaizon, J.-F. Humbert, et al. 2017. "Overview of Freshwater Microbial Eukaryotes Diversity: A First Analysis of Publicly Available Metabarcoding Data." *FEMS Microbiology Ecology* 93, no. 4: fix023. <https://doi.org/10.1093/femsec/fix023>.

Del Cortona, A., F. Leliaert, K. A. Bogaert, et al. 2017. "The Plastid Genome in Cladophorales Green Algae Is Encoded by Hairpin Chromosomes." *Current Biology* 27, no. 24: 3771–3782.e6. <https://doi.org/10.1016/j.cub.2017.11.004>.

Delmont, T. O., M. Gaia, D. D. Hinsinger, et al. 2022. "Functional Repertoire Convergence of Distantly Related Eukaryotic Plankton Lineages Abundant in the Sunlit Ocean." *Cell Genomics* 2, no. 5: 100123. <https://doi.org/10.1016/j.xgen.2022.100123>.

Diaz, J. M., and S. Plummer. 2018. "Production of Extracellular Reactive Oxygen Species by Phytoplankton: Past and Future Directions." *Journal of Plankton Research* 40, no. 6: 655–666. <https://doi.org/10.1093/plankt/fby039>.

Duncan, A., K. Barry, C. Daum, et al. 2022. "Metagenome-Assembled Genomes of Phytoplankton Microbiomes From the Arctic and Atlantic Oceans." *Microbiome* 10, no. 1: 67. <https://doi.org/10.1186/s40168-022-01254-7>.

Eme, L., and D. Tamarit. 2024. "Microbial Diversity and Open Questions About the Deep Tree of Life." *Genome Biology and Evolution* 16, no. 4: evae053. <https://doi.org/10.1093/gbe/evae053>.

Eren, A. M., Ö. C. Esen, C. Quince, et al. 2015. "Anvi'o: An Advanced Analysis and Visualization Platform for 'Omics Data." *PeerJ* 3: e1319. <https://doi.org/10.7717/peerj.1319>.

Fang, L., F. Leliaert, Z. Zhang, D. Penny, and B. Zhong. 2017. "Evolution of the Chlorophyta: Insights From Chloroplast Phylogenomic Analyses." *Journal of Systematics and Evolution* 55, no. 4: 322–332. <https://doi.org/10.1111/jse.12248>.

Fritz, A., P. Hofmann, S. Majda, et al. 2019. "CAMISIM: Simulating Metagenomes and Microbial Communities." *Microbiome* 7, no. 1: 17. <https://doi.org/10.1186/s40168-019-0633-6>.

Gallagher, S. D., S. T. Fitz-Gibbon, D. Strenkert, S. O. Purvine, M. Pellegrini, and S. S. Merchant. 2018. "High-Throughput Sequencing of the Chloroplast and Mitochondrion of *Chlamydomonas reinhardtii* to Generate Improved De Novo Assemblies, Analyze Expression Patterns and Transcript Speciation, and Evaluate Diversity Among Laboratory Strains and Wild Isolates." *Plant Journal* 93, no. 3: 545–565. <https://doi.org/10.1111/tpj.13788>.

Gao, X., K. Chen, J. Xiong, et al. 2024. "The P10K Database: A Data Portal for the Protist 10 000 Genomes Project." *Nucleic Acids Research* 52, no. D1: D747–D755. <https://doi.org/10.1093/nar/gkad992>.

GenBank. 2025. "National Center for Biotechnology Information." <https://www.ncbi.nlm.nih.gov/>.

- Guiry, M. D. 2024. "How Many Species of Algae Are There? A Reprise. Four Kingdoms, 14 Phyla, 63 Classes and Still Growing." *Journal of Phycology* 60, no. 2: 214–228. <https://doi.org/10.1111/jpy.13431>.
- Hickl, O., P. Queirós, P. Wilmes, P. May, and A. Heintz-Buschart. 2022. "Binny: An Automated Binning Algorithm to Recover High-Quality Genomes From Complex Metagenomic Datasets." *Briefings in Bioinformatics* 23, no. 6: bbac431. <https://doi.org/10.1093/bib/bbac431>.
- Howe, C. J., R. E. R. Nisbet, and A. C. Barbrook. 2008. "The Remarkable Chloroplast Genome of Dinoflagellates." *Journal of Experimental Botany* 59, no. 5: 1035–1045. <https://doi.org/10.1093/jxb/erm292>.
- Huang, L., H. Yu, Z. Wang, and W. Xu. 2024. "CPStools: A Package for Analyzing Chloroplast Genome Sequences." *iMetaOmics* 1, no. 2: e25. <https://doi.org/10.1002/imo2.25>.
- Hug, L. A., B. J. Baker, K. Anantharaman, et al. 2016. "A New View of the Tree of Life." *Nature Microbiology* 1, no. 5: 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>.
- Jamy, M., T. Huber, T. Antoine, et al. 2025. "Identification of a Deep-Branching Lineage of Algae Using Environmental Plastid Genomes." *Nature Communications* 17, no. 1: 662. <https://doi.org/10.1038/s41467-025-67401-4>.
- Kang, D. D., F. Li, E. Kirton, et al. 2019. "MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction From Metagenome Assemblies." *PeerJ* 7: e7359. <https://doi.org/10.7717/peerj.7359>.
- Karlicki, M. 2024. *Diversity and Ecology of Photosynthetic Microbial Eukaryotes in Selected Aquatic Systems Based on Metabarcoding and Metagenomic Data*. Institutional Repository of the University of Warsaw. <https://repozytorium.uw.edu.pl/bitstreams/0ac4dc59-9146-4b3a-a8db-3f0f84947dca/download>.
- Karlicki, M., S. Antonowicz, and A. Karnkowska. 2022. "Tiara: Deep Learning-Based Classification System for Eukaryotic Sequences." *Bioinformatics* 38, no. 2: 344–350. <https://doi.org/10.1093/bioinformatics/btab672>.
- Krinos, A. I., R. M. Bowers, R. R. Rohwer, K. D. McMahon, T. Woyke, and F. Schulz. 2024. "Time-Series Metagenomics Reveals Changing Protistan Ecology of a Temperate Dimictic Lake." *Microbiome* 12, no. 1: 133. <https://doi.org/10.1186/s40168-024-01831-y>.
- Kutuzova, S., M. Nielsen, P. Piera, J. N. Nissen, and S. Rasmussen. 2024. "Taxometer: Improving Taxonomic Classification of Metagenomics Contigs." *Nature Communications* 15, no. 1: 8357. <https://doi.org/10.1038/s41467-024-52771-y>.
- Laforest-Lapointe, I., and M.-C. Arrieta. 2018. "Microbial Eukaryotes: A Missing Link in Gut Microbiome Studies." *mSystems* 3, no. 2: e00201-17. <https://doi.org/10.1128/mSystems.00201-17>.
- Leão, P., M. E. Little, K. E. Appler, et al. 2024. "Asgard Archaea Defense Systems and Their Roles in the Origin of Eukaryotic Immunity." *Nature Communications* 15, no. 1: 6386. <https://doi.org/10.1038/s41467-024-50195-2>.
- Li, H. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." *Bioinformatics* 34, no. 18: 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- Liu, Y., K. S. Makarova, W.-C. Huang, et al. 2021. "Expanded Diversity of Asgard Archaea and Their Relationships With Eukaryotes." *Nature* 593, no. 7860: 553–557. <https://doi.org/10.1038/s41586-021-03494-3>.
- Lopes Dos Santos, A., P. Gourvil, M. Tragin, et al. 2017. "Diversity and Oceanic Distribution of Prasinophytes Clade VII, the Dominant Group of Green Algae in Oceanic Waters." *ISME Journal* 11, no. 2: 512–528. <https://doi.org/10.1038/ismej.2016.120>.
- Marcelino, V. R., P. T. L. C. Clausen, J. P. Buchmann, et al. 2020. "CCMetagen: Comprehensive and Accurate Identification of Eukaryotes and Prokaryotes in Metagenomic Data." *Genome Biology* 21, no. 1: 103. <https://doi.org/10.1186/s13059-020-02014-2>.
- Meyer, F., P. Hofmann, P. Belmann, et al. 2018. "Amber: Assessment of Metagenome Binners." *GigaScience* 7, no. 6: giy069. <https://doi.org/10.1093/gigascience/giy069>.
- Miao, W., L. Song, S. Ba, et al. 2020. "Protist 10,000 Genomes Project." *Innovation* 1, no. 3: 100058. <https://doi.org/10.1016/j.xinn.2020.100058>.
- Mölder, F., K. P. Jablonski, B. Letcher, et al. 2021. "Sustainable Data Analysis With Snakemake." *F1000Research* 10: 33. <https://doi.org/10.12688/f1000research.29032.1>.
- Nayfach, S., S. Roux, R. Seshadri, et al. 2021. "A Genomic Catalog of Earth's Microbiomes." *Nature Biotechnology* 39, no. 4: 499–509. <https://doi.org/10.1038/s41587-020-0718-6>.
- O'Leary, N. A., M. W. Wright, J. R. Brister, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44, no. D1: D733–D745. <https://doi.org/10.1093/nar/gkv1189>.
- Ondov, B. D., N. H. Bergman, and A. M. Phillippy. 2011. "Interactive Metagenomic Visualization in a Web Browser." *BMC Bioinformatics* 12, no. 1: 385. <https://doi.org/10.1186/1471-2105-12-385>.
- Pan, S., X.-M. Zhao, and L. P. Coelho. 2023. "SemiBin2: Self-Supervised Contrastive Learning Leads to Better MAGs for Short- and Long-Read Sequencing." *Bioinformatics* 39, no. Supplement_1: i21–i29. <https://doi.org/10.1093/bioinformatics/btad209>.
- Parks, D. H., M. Chuvpochina, C. Rinke, A. J. Mussig, P.-A. Chaumeil, and P. Hugenholtz. 2022. "GTDB: An Ongoing Census of Bacterial and Archaeal Diversity Through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy." *Nucleic Acids Research* 50, no. D1: D785–D794. <https://doi.org/10.1093/nar/gkab776>.
- Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. 2015. "CheckM: Assessing the Quality of Microbial Genomes Recovered From Isolates, Single Cells, and Metagenomes." *Genome Research* 25, no. 7: 1043–1055. <https://doi.org/10.1101/gr.186072.114>.
- Patin, N. V., and K. D. Goodwin. 2022. "Long-Read Sequencing Improves Recovery of Picoeukaryotic Genomes and Zooplankton Marker Genes From Marine Metagenomes." *mSystems* 7, no. 6: e00595-22. <https://doi.org/10.1128/msystems.00595-22>.
- Penot, M., J. B. Dacks, B. Read, and R. G. Dorrell. 2022. "Genomic and Meta-Genomic Insights Into the Functions, Diversity and Global Distribution of Haptophyte Algae." *Applied Phycology* 3, no. 1: 340–359. <https://doi.org/10.1080/26388081.2022.2103732>.
- Pierella Karlusich, J. J., E. Pelletier, L. Zinger, et al. 2023. "A Robust Approach to Estimate Relative Phytoplankton Cell Abundances From Metagenomes." *Molecular Ecology Resources* 23, no. 1: 16–40. <https://doi.org/10.1111/1755-0998.13592>.
- Piganeau, G., and H. Moreau. 2007. "Screening the Sargasso Sea Metagenome for Data to Investigate Genome Evolution in *Ostreococcus* (Prasinophyceae, Chlorophyta)." *Gene* 406, no. 1–2: 184–190. <https://doi.org/10.1016/j.gene.2007.09.015>.
- Quince, C., A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. 2017. "Shotgun Metagenomics, From Sampling to Analysis." *Nature Biotechnology* 35, no. 9: 833–844. <https://doi.org/10.1038/nbt.3935>.
- Salazar, V. W., B. Shaban, M. D. M. Quiroga, et al. 2022. "Metaphor—A Workflow for Streamlined Assembly and Binning of Metagenomes." *GigaScience* 12: giad055. <https://doi.org/10.1093/gigascience/giad055>.
- Sato, S. 2011. "The Apicomplexan Plastid and Its Evolution." *Cellular and Molecular Life Sciences* 68, no. 8: 1285–1296. <https://doi.org/10.1007/s00018-011-0646-1>.
- Sauvage, T., W. E. Schmidt, S. Suda, and S. Fredericq. 2016. "A Metabarcoding Framework for Facilitated Survey of Endolithic Phototrophs With tufA." *BMC Ecology* 16, no. 1: 8. <https://doi.org/10.1186/s12898-016-0068-x>.

- Schmidt, T. S. B., A. Fullam, P. Ferretti, et al. 2024. "Spire: A Searchable, Planetary-Scale Microbiome Resource." *Nucleic Acids Research* 52, no. D1: D777–D783. <https://doi.org/10.1093/nar/gkad943>.
- Sibbald, S. J., and J. M. Archibald. 2017. "More Protist Genomes Needed." *Nature Ecology & Evolution* 1, no. 5: 0145. <https://doi.org/10.1038/s41559-017-0145>.
- Sibbald, S. J., and J. M. Archibald. 2020. "Genomic Insights Into Plastid Evolution." *Genome Biology and Evolution* 12, no. 7: 978–990. <https://doi.org/10.1093/gbe/evaa096>.
- Solomon, R., T. Wein, B. Levy, et al. 2022. "Protozoa Populations Are Ecosystem Engineers That Shape Prokaryotic Community Structure and Function of the Rumen Microbial Ecosystem." *ISME Journal* 16, no. 4: 1187–1197. <https://doi.org/10.1038/s41396-021-01170-y>.
- Steenwyk, J. L., and A. Rokas. 2021. "Orthofisher: A Broadly Applicable Tool for Automated Gene Identification and Retrieval." *G3: Genes, Genomes, Genetics* 11, no. 9: jkab250. <https://doi.org/10.1093/g3journal/jkab250>.
- Sun, L., L. Fang, Z. Zhang, X. Chang, D. Penny, and B. Zhong. 2016. "Chloroplast Phylogenomic Inference of Green Algae Relationships." *Scientific Reports* 6, no. 1: 20528. <https://doi.org/10.1038/srep20528>.
- Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. 2007. "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters." *Bioinformatics* 23, no. 10: 1282–1288. <https://doi.org/10.1093/bioinformatics/btm098>.
- Terpis, K. X., E. D. Salomaki, D. Barcytė, et al. 2025. "Multiple Plastid Losses Within Photosynthetic Stramenopiles Revealed by Comprehensive Phylogenomics." *Current Biology* 35, no. 3: 483–499.e8. <https://doi.org/10.1016/j.cub.2024.11.065>.
- The Genome Standards Consortium, R. M. Bowers, N. C. Kyrpides, et al. 2017. "Minimum Information About a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35, no. 8: 725–731. <https://doi.org/10.1038/nbt.3893>.
- Turmel, M., C. Otis, and C. Lemieux. 2015. "Dynamic Evolution of the Chloroplast Genome in the Green Algal Classes Pedinophyceae and Trebouxiophyceae." *Genome Biology and Evolution* 7, no. 7: 2062–2082. <https://doi.org/10.1093/gbe/evv130>.
- Turnbull, R. 2025. "Rbturnbull/Corgi [GitHub Repository]." *GitHub*. <https://github.com/rbturnbull/Corgi>.
- Van der Jeugt, F., P. Dawyndt, and B. Mesuere. 2022. "FragGeneScanRs: Faster Gene Prediction for Short Reads." *BMC Bioinformatics* 23, no. 1: 198. <https://doi.org/10.1186/s12859-022-04736-5>.
- Von Meijenfeldt, F. A. B., K. Arkhipova, D. D. Cambuy, F. H. Coutinho, and B. E. Dutilh. 2019. "Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins With CAT and BAT." *Genome Biology* 20, no. 1: 217. <https://doi.org/10.1186/s13059-019-1817-x>.
- Wang, Z., R. You, H. Han, W. Liu, F. Sun, and S. Zhu. 2024. "Effective Binning of Metagenomic Contigs Using Contrastive Multi-View Representation Learning." *Nature Communications* 15, no. 1: 585. <https://doi.org/10.1038/s41467-023-44290-z>.
- Warnock, J. P., R. P. Scherer, and M. A. Konfirst. 2015. "A Record of Pleistocene Diatom Preservation From the Amundsen Sea, West Antarctica With Possible Implications on Silica Leakage." *Marine Micropaleontology* 117: 40–46. <https://doi.org/10.1016/j.marmicro.2015.04.001>.
- Wideman, J. G., A. Monier, R. Rodríguez-Martínez, et al. 2019. "Unexpected Mitochondrial Genome Diversity Revealed by Targeted Single-Cell Genomics of Heterotrophic Flagellated Protists." *Nature Microbiology* 5, no. 1: 154–165. <https://doi.org/10.1038/s41564-019-0605-4>.
- Wirth, W., S. Mutch, and R. Turnbull. 2024. "Snk: A Snakemake CLI and Workflow Management System." *Journal of Open Source Software* 9, no. 103: 7410. <https://doi.org/10.21105/joss.07410>.
- Xiao, X., W. Zhao, Z. Song, et al. 2025. "Microbial Ecosystems and Ecological Driving Forces in the Deepest Ocean Sediments." *Cell* 188, no. 5: 1363–1377.e9. <https://doi.org/10.1016/j.cell.2024.12.036>.
- Yang, Y., J. Sun, C. Chen, et al. 2022. "Metagenomic and Metatranscriptomic Analyses Reveal Minor-Yet-Crucial Roles of Gut Microbiome in Deep-Sea Hydrothermal Vent Snail." *Animal Microbiome* 4, no. 1: 3. <https://doi.org/10.1186/s42523-021-00150-z>.
- Zhang, M., Z. Cui, F. Liu, and N. Chen. 2021. "Complete Chloroplast Genome of *Eucampia zodiacus* (Mediophyceae, Bacillariophyta)." *Mitochondrial DNA Part B Resources* 6, no. 8: 2194–2197. <https://doi.org/10.1080/23802359.2021.1944828>.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1:** Top 10 BLASTN search hits (with core nucleotide (nt) database) for each bin's *rbcl* nucleotide sequence for sample SAMEA2189670 and SAMEA2732360. **File S1:** genome id, count of contigs, percentage of genome completeness, name, GenBank ID, relative abundance of the chosen genomes in two simulated samples used for benchmarking. **Table S1:** Amber results comparing MetaBAT2 with binny from ChloroScan using single sample synthetic metagenome 1. Each row metric definition can be found in the definition of terms by AMBER above. **Table S2:** Amber results comparing MetaBAT2 with binny from ChloroScan using single sample synthetic metagenome 2. Each row's metric definition can be found in the definition of terms by AMBER above. **Figure S1:** Amber results showing the benchmarked binning tools' efforts, showing the adjusted random Index vs. Percentage of binned contigs (bp) of binny and MetaBAT2 from (a) the synthetic single-sample metagenome 1 and (b) the single-sample synthetic metagenome 2. (c) The average completeness (%) vs. Average purity (%) of MAGs recovered by binny and MetaBAT2 from the single-sample synthetic metagenome 1; (d) The average completeness (%) vs. Average purity (%) of MAGs recovered by binny and MetaBAT2 from the single-sample synthetic metagenome 2. **Figure S2:** The GC content vs. Log read depth scatterplot of (a) synthetic single sample metagenome 1 and (b) synthetic single sample metagenome 2. Each dot represents a contig and their size is scaled by the number of marker genes annotated on them; larger dots refer to contigs with more marker genes. **Figure S3:** Percentages of count of contigs from each source genome out of all contigs from the plastid-enriched assembly in two benchmarking samples. Prokaryotic sequences are coloured black. This suggests a potential inclusion of prokaryotic and nuclear genome contaminations by Corgi. **Figure S4:** Scatterplot (a) and taxonomic classification (b) of two bins from SAMEA2189670. **Figure S5:** Krona plots show the count of eukaryote contigs for each taxon from the analysed Tara Oceans samples. (a) SAMEA2189670; (b) SAMEA2732613; (c) SAMEA2657032 and (d) SAMEA2732360. Percentages are not scaled by the abundance of contigs, but only the count of contigs belonging to the corresponding taxon.