

Taxonomy bias in metagenome-assembled genome recovery

Daniel J. Nebauer^{1,2}, Tiffanie Nelson^{1,3}, Caitlin Romanis¹, Brett A. Neilan^{1,2,*} and Verlaine J. Timms^{1,2}

Abstract

The recovery of metagenome-assembled genomes (MAGs) from shotgun metagenomic sequencing is rapidly expanding the availability of representative genomes. However, this practice may skew the representation of specific taxa in real-world datasets. This bias is attributed primarily to the known inefficiencies of sequence-by-synthesis platforms in amplifying GC-rich and AT-rich sequence fragments. Here, we recover 216 medium- and high-quality MAGs from an Australian wetland site. Notably, no MAGs were recovered for some dominant cyanobacterial and proteobacterial species known to be present. A new protocol involving read-based classification and alignment to the MAG dataset demonstrated the highly efficient recovery of low-GC organisms in the *Actinobacteria* and *Bacteroidota* phyla. Additionally, the recovery of lost taxonomic information was demonstrated through unmatched sample mapping. The findings suggest a bias towards the recovery of smaller, low-GC organisms in MAG recovery, potentially skewing the global representation of microbial diversity. Our pipeline is made publicly available as a tool to help researchers estimate taxonomic losses following MAG recovery efforts.

Impact Statement

Metagenome-assembled genomes (MAGs) have become central to genome-resolved microbial ecology, yet the potential biases introduced by assembly and binning remain underexplored in real-world datasets. This study demonstrates that MAG recovery does not necessarily reflect organismal abundance, revealing systematic taxonomic distortions linked to GC content and genome size. Using shotgun metagenomes from an Australian wetland, we show preferential recovery of smaller, low-GC genomes and the loss of dominant taxa, including ecologically significant *Cyanobacteria*. We further introduce a read-based mapping strategy that partially restores lost taxonomic information, providing a practical framework for assessing MAG recovery efficiency. These findings highlight an important limitation of current MAG-centric analyses and offer methodological guidance to improve the interpretation of genome-resolved metagenomic studies.

DATA AVAILABILITY

All raw shotgun metagenomic sequencing data generated in this study are publicly available through the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1230272.

The ten individual SRA accessions are SRR32538009–SRR32538018.

Scripts for data analysis and visualization are publicly available in the mag-attribution-tool repository: <https://github.com/DanielNebauer94/mag-attribution-tool>.

Received 03 March 2026; Accepted 12 May 2026; Published 08 June 2026

Author affiliations: ¹School of Environmental and Life Sciences, The University of Newcastle, Callaghan, New South Wales, Australia; ²Australian Research Council Centre of Excellence in Synthetic Biology, Macquarie Park, NSW, 2109, Australia; ³Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Victoria, Australia.

***Correspondence:** Brett A. Neilan, Brett.neilan@newcastle.edu.au

Keywords: GC bias; metagenome-assembled genomes (MAGs); microbial diversity; shotgun metagenomics; taxonomic bias.

Abbreviations: MAG, metagenome-assembled genome.

All supporting data, code and protocols have been provided within the article or through supplementary data files. Nine supplementary data are available with the online version of this article.

001736 © 2026 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

INTRODUCTION

Shotgun metagenomics is an established but evolving field enabling the genomic profiling of microbial communities. It increases taxonomic depth and resolution over 16S rRNA amplicon sequencing approaches [1], while also providing insights into community functions [2]. Of increasing importance in microbial ecology is the recovery of metagenome-assembled genomes (MAGs), which are reconstructions of complete or near-complete genomes from shotgun metagenomic data [3]. MAG recovery has revolutionized the discovery of a vast number of novel microbes which are not amenable to cultivation and has seen a substantial expansion of the tree of life [4, 5]. Further, MAG recovery circumvents time-consuming isolation and sequencing of single organisms, enabling genomic profiling of microbes often at the genus or species level, and in some instances may discriminate between strains or ecotypes [6].

MAG recovery requires four essential steps: (1) DNA from a microbial community is sequenced into reads, (2) reads are assembled into longer contiguous sequences called contigs, (3) contigs are grouped together based on sequence similarity, a practice called binning, and (4) bins are assessed for completion and contamination and taxonomy is inferred through phylogenetic placement into the tree of life [7]. Bioinformatic pipelines to bin contigs and ease MAG recovery are rapidly evolving [8]; however, most current algorithms bin contigs based on a combination of tetranucleotide frequencies and contig coverage values [9].

As MAGs rapidly expand the global catalogue of microbial genomes [4], it is becoming increasingly important to consider how this technology may be skewing the representation of specific taxa. The GC content of bacterial genomes can vary widely and significantly affect sequencing and assembly efforts [10, 11]. High-throughput sequencing platforms which employ sequence by synthesis technologies [12] are known to inefficiently amplify both GC and AT-rich sequence fragments [13, 14]. Reduced read coverage of GC-extreme regions results in fragmented assemblies, hindering the recovery of high-quality genomes from organisms with extreme GC content [15]. In the context of MAG recovery, fragmented assemblies reduce the genomic template available for robust assessment of tetranucleotide frequencies during binning [16], theoretically disfavoring recovery of genomes with extreme GC content at the sequencing and assembly stages. While GC bias is well-documented in the recovery of GC-extreme microbes and mock community datasets [15, 17], little attention has been given to its potential consequences in real-world metagenomic datasets. Further, very few bioinformatic tools exist to examine this problem in a straightforward manner.

Here, we employ a widely adopted pipeline to recover MAGs from shotgun metagenomic sequencing data generated in a real-world, year-long study investigating a wetland site in Australia. The site experienced high season-to-season taxonomic variability, with one sample dominated by *Planktothrix*, an AT-rich genus of *Cyanobacteria* [18]. We identify taxonomic information lost during MAG recovery and investigate the role of GC content. We also demonstrate the use of read alignment to unmatched samples as a viable way to recover taxonomic information lost during MAG recovery. We make this pipeline publicly available as a practical tool for researchers to quantify taxonomic attrition following MAG recovery efforts, deployable on modest computational infrastructure without the need for high-performance computing resources.

METHODS

Sample collection

Water samples were collected according to the sample key in Table 1. Sampling sites were termed ‘Inlet’ due to the proximity to an inlet pipe and ‘Colony’ due to the proximity to a flying fox colony. Water samples were collected with a sterilized polypropylene bottle. The bottle was washed three times with water close to the sample site prior to sample collection and care was taken not to disturb the water column. Biomass was filtered onto 1.2 µm glass fibre filters within 5 h of collection. Filters containing biomass were stored at −30 °C in 1.5 ml Eppendorf tubes until required.

Genomic DNA extraction, sequencing and quality control

Genomic DNA extraction from filters was performed using the FastDNA Spin Kit for Soil (MP Bio, Solon, Ohio), following the manufacturer’s instructions. DNA was quantified using the high-sensitivity Qubit dsDNA assay (Invitrogen). Paired-end sequencing and demultiplexing was performed at the Ramaciotti Genomic Research Centre (UNSW) using a NovaSeq (Illumina) with a NovaSeq 6000 SP reagent kit (300 cycles) at a read length of 150 bp. Quality trimming and the removal of adapters and unpaired reads were performed using Trimmomatic v0.32 [19], with the following arguments: -phred33 -leading:3 -trailing:3 -slidingwindow:4:15 -minlen:36. Read quality in each sample was assessed with FastQC (v0.12.1) and visualized with MultiQC (v1.16), with phred scores >20 used as the metric for quality passing, which allows 1% base call error [20].

Assembly, binning, MAG recovery and relative abundance

All analyses were performed on a workstation running Windows 10 with the Windows Subsystem for Linux (WSL), equipped with an Intel Core i5-12400F processor, 128 GB DDR4 RAM, and a 3 TB external solid state drive.

Table 1. Sample key. A total of 10 samples underwent shotgun metagenomic sequencing. Samples encompassed two sites (Inlet and Colony) and five dates, which spanned 11 months

Sample name	Site	Date
RW01	Inlet	16/06/2020
RW02	Inlet	08/09/2020
RW03	Inlet	29/12/2020
RW04	Inlet	23/02/2021
RW05	Inlet	08/04/2021
RW06	Colony	16/06/2020
RW07	Colony	08/09/2020
RW08	Colony	29/12/2020
RW09	Colony	23/02/2021
RW10	Colony	08/04/2021

The appropriate algorithm for assembly and binning depends on the dataset [21, 22], so two methods for assembly and three methods for binning were compared. High-quality reads were assembled into contigs/scaffolds with metaSPAdes v3.15.2 and MEGAHIT v1.1.4. A co-assembly using either the MEGAHIT or metaSPAdes assemblers was not possible using the available computational infrastructure [23], so each dataset was assembled individually. MetaQUAST v5.2.0 [24] was run in reference-free mode to assess assembly quality.

Contigs underwent binning via the metabat2 v2.15 [25], maxbin2 v2.2.7 [26] and metaWRAP v1.3 pipelines, with minimum contig length set at 1,000 bp. Bin completion and contamination metrics were computed using checkM v1.2.2 and a set of 43 phylogenetically informative marker genes [27]. This study followed minimum information about a MAG standards to define genome quality [28]; $\geq 50\%$ completion and $\leq 10\%$ contamination were defined as medium quality, and $\geq 90\%$ completion and $\leq 5\%$ contamination were defined as high quality. Low-quality MAGs ($\leq 50\%$ completion and/or $\geq 10\%$ contamination) were removed. Bins were classified with GTDB-Tk v2.3.2 through placement into a reference tree against genomes from the GTDB database (release 207) [7]. At this point, they were defined as MAGs.

MAG relative abundance was computed using coverM v0.6.1 (<https://github.com/wwood/CoverM>). Contig edges were trimmed by 75 bp to avoid contig edge coverage bias [29], and MAGs with less than 10% total coverage were not considered. Relative abundance for each MAG was calculated as the number of reads aligned to a MAG divided by the total number of bacterial reads in each metagenomic dataset.

Phylogenetic tree construction

A phylogenetic tree was constructed using GToTree v1.8.1 [30] and the gene set (HMM file) for bacteria (74 targets). Coding sequences were retrieved using Prodigal v2.6.3 [31], and HMMER3 v3.2.1 [32] was used to identify target genes present in the HMM file. Gene alignments were then performed by MUSCLE v5.1 [33], and alignments were trimmed using Trimal v1.4 [34]. The concatenated aligned genes were then used to construct an approximate maximum likelihood tree using FastTree 2 v2.1.11 [35]. The tree was visualized and decorated using iTOL v6.9 [36].

Taxa recovery estimations

High-quality reads were taxonomically classified with Kraken2 v2.1.1 [37], using a 64 GB kraken-compatible version of the GTDB database (created 23 July 2019) [38]; as such, nomenclature used throughout this manuscript follows GTDB conventions. Read classification results from the kraken2 kreport file were computed and visualized in R v4.2.0 with the Pavian package v1.2.1 [39].

To investigate read recovery into MAGs, bacterial reads were first extracted from fastq files with classification information obtained from kraken2 files. Extracted reads were then aligned to MAG datasets using BWA-MEM v0.7.18 [40] with default parameters. Samtools v1.2.0 [41] was used to convert resultant sequence alignment map (SAM) files to binary (BAM) format and to isolate aligned and unaligned reads into separate fastq files. For each genus in each sample, an in-house python script was used to determine the proportion of aligned vs unaligned reads and the average read GC content.

Barplots, density plots, bubble plots and heatmaps were generated in R v4.2.0 using the ggplot2 v3.5.1 [42] and pheatmap v1.0.12 [43] packages. Bubble plots were generated using a custom python script publicly available at <https://github.com/DanielNebauer94/mag-attribution-tool>.

RESULTS

Read-based community composition

We analysed ten metagenomic samples (RW01–RW10) collected from the Inlet and Colony sites across five seasonal time points. Reads were taxonomically classified using Kraken2 against a GTDB-compatible reference database (see methods). An average number of 42.9 million reads were obtained from all samples, ranging from 32.7 to 48.5 million. Of these, an average of 22.17% were classified as bacteria and archaea, with the highest percentage of read classification achieved during late summer at the Inlet site (RW09; 38.5%) and the lowest during autumn at the Colony site (RW10; 6.77%). Additional information about read quality and classification is available in the supplementary (Data S1 and S2, available in the online Supplementary Material). Read classification yielded 151 phyla, including 16 archaea. As this study focused on the bacterial community, non-bacterial reads were excluded from downstream analysis. Of classified bacterial reads, samples at both sites predominantly consisted of *Proteobacteria* (59%), *Cyanobacteria* (11%), *Actinobacteria* (11%) and *Bacteroidota* (9%) (Fig. 1). The relative abundance of *Cyanobacteria*, *Bacteroidota* and *Proteobacteria* was highly variable across samples, as were major genera such as *Planktothrix*, *Limnohabitans*, *Flavobacterium* and *Microcystis*.

The percentage of reads classified as *Cyanobacteria* in the water samples ranged from 0.33 to 53.15% with the highest abundance occurring during late summer at the Inlet site (sample RW04). This sample was dominated by the genus *Planktothrix*, which comprised 41.76% of the total microbial community. Meanwhile, the Colony site experienced low cyanobacterial abundance (2.53%), though *Planktothrix* still dominated the cyanobacterial community (32.13%). Cyanobacterial abundance decreased sharply at the Inlet site during autumn (1.97%), though *Planktothrix* and *Microcystis* remained the dominant members of the cyanobacterial community (19.95 and 18.92%, respectively). Meanwhile, cyanobacterial populations grew to 16.82% at the Colony

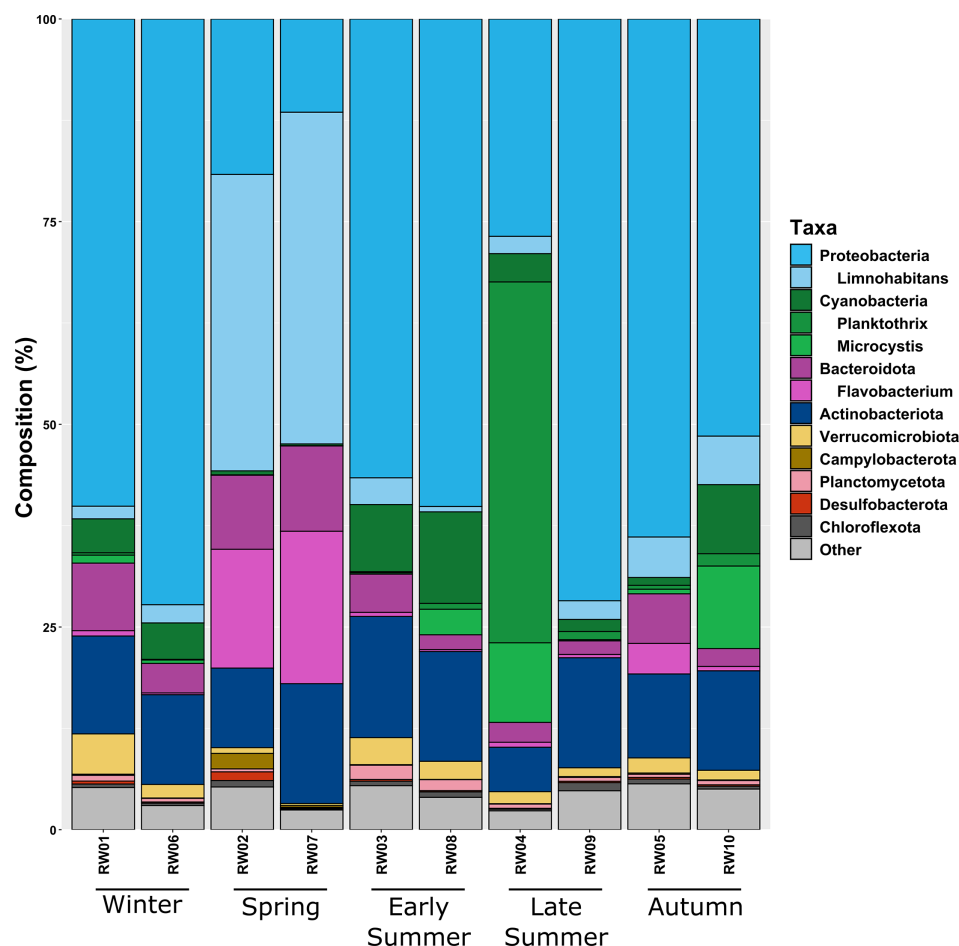


Fig. 1. Bacterial community abundance. Bacterial abundance was determined according to metagenomic read-based taxonomic classification. Only reads classified as bacterial were considered, and thus the bar plot represents the relative abundance of each taxon as a proportion of classified bacterial reads only. Samples are clustered according to season. Phyla with >1% relative composition in at least one sample are presented. Also presented are genera with high abundance (>10%) in at least one sample. These genera are indented under their respective phyla in the taxa key.

site and the cyanobacterial community became dominated by *Microcystis* (38.52%), while *Planktothrix* populations decreased (7.09%).

In early summer, at the Inlet site, the cyanobacterial community was dominated by microbes from the *Synechococcaceae* family, the most abundant genus being *Vulcanococcus* (38.92%). *Microcystis* populations made up 14.50% of the cyanobacterial community at the Colony site. Winter saw a richness in diazotrophic *Nodularia* species, which comprised 20.52 and 36.25% of the cyanobacterial community at the Inlet and Colony sites, respectively.

Spring saw negligible cyanobacterial populations at both sites; 0.63% at the Inlet and 0.32% at the Colony sites. This season instead showed domination by heterotrophic *Limnohabitans* from the *Proteobacteria* phylum, which comprised 28.68 and 34.58% of the Inlet and Colony populations, respectively, and *Flavobacterium* from the *Bacteroidota* phylum, which comprised 13.29 and 18.17% at the Inlet and Colony sites, respectively.

MAG dataset

Assembly and binning were undertaken separately for each sample. Two assemblers and three binning algorithms were compared to generate the highest-quality dataset which captured the most bins. Comparison of two assemblers (metaSPAdes [44] and MEGAHIT [45]) found metaSPAdes total contig length was superior to that of MEGAHIT. MetaSPAdes also produced the most contigs $\geq 50,000$ and $\geq 1,000$ bp; however, MEGAHIT produced the most contigs $\geq 25,000$ bp, $\geq 10,000$ bp and $\geq 5,000$ bp. MEGAHIT also produced contigs with a greater N50 value in all samples, except RW10 (Data S3). Use of the metaWRAP [46] binning tool resulted in more bins than metatb2 and maxbin2 in every sample and produced more high-quality and marginally less medium-quality bins when supplied with MEGAHIT assemblies. Read assembly using MEGAHIT and contig binning using the metaWRAP tool was therefore pursued in the generation of the MAG dataset.

A total of 216 MAGs were recovered from the 10 shotgun metagenome datasets (Data S5). Of these, 62 were defined as high quality, and 154 defined as medium quality. The total MAG dataset comprised 5.76×10^8 nucleotides and captured a total of 14 phyla. Most recovered MAGs were classified as *Proteobacteria* (37%), followed by *Bacteroidota* (22%), *Actinobacteriota* (15%) and *Cyanobacteria* (8%). Median completion scores for MAGs in most phyla were $\geq 70\%$, except for *Bdellovibrionota* (58%) and

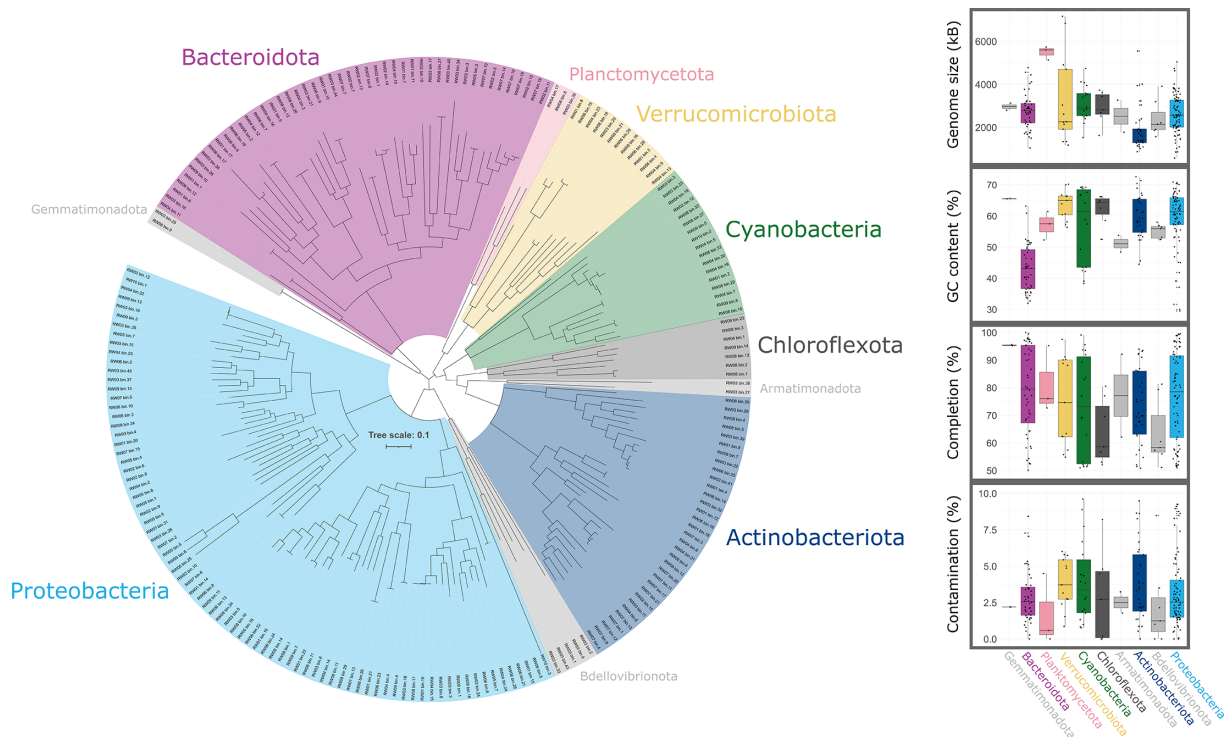


Fig. 2. Phylogenetic tree and basic statistics for the MAG dataset. The approximate maximum-likelihood tree was inferred from the concatenated alignment of 74 proteins using FastTree and the GToTree workflow. Branches and clades are coloured and labelled based on the taxonomic annotations estimated by GTDB-Tk. Boxplots showing the distributions of MAG genome size (kB), completeness, contamination and GC content scores are coloured to match the phylogenetic tree. The interactive tree may be accessed here: <https://itol.embl.de/tree/1341488223716411687316963>.

Chloroflexota (58%), while median contamination scores in all phyla did not exceed 5%. *Bacteroidota* had the lowest median score for GC content (43%), while *Cyanobacteria* had the greatest GC content interquartile range (Fig. 2).

Taxonomic recovery and GC content effects

To assess taxonomy lost during assembly and binning, bacterial reads were aligned to the MAG datasets. An average of 12.3% of reads aligned to the MAG datasets, ranging from 1% (sample RW05) to 25% (sample RW07). It is acknowledged that this approach captures only the intersection of two independently biased analyses, that is, read-based taxonomic classification and MAG recovery, and that unclassified reads may also map to the MAG dataset. However, total read-to-MAG mapping was independently assessed via CoverM during relative abundance estimation (Data S5), providing a complementary community-level view that is not limited by classification rate. Most lost information belonged to *Proteobacteria* in almost all samples (Data S6), as this phylum constituted the majority of taxonomic classifications and the highest intra-phyla diversity at the species-level, which complicates high-quality MAG recovery. One sample, RW04, was an exception, where information lost belonged to *Cyanobacteria*.

Most samples had relatively high binning efficiencies for *Bacteroidota* (Fig. 3), particularly during spring where this phylum was most abundant (samples RW02 and RW07). Binning efficiency here refers to the proportion of reads classified to a given phylum that could be successfully aligned to a MAG from that phylum; a value of 100% would indicate complete recovery of all reads from that taxon into the MAG dataset, while 0% indicates no recovery. Binning efficiencies for *Actinobacteriota* were also notably high in these samples (33 and 66.4% in samples RW02 and RW07, respectively). A higher binning efficiency for cyanobacteria was noted in sample RW09 (49%). Surprisingly, a *Planktothrix* MAG was recovered in this sample (Data S5; RW09_bin.8), though not in the sample where *Planktothrix* was far more abundant (Fig. 1; sample RW04).

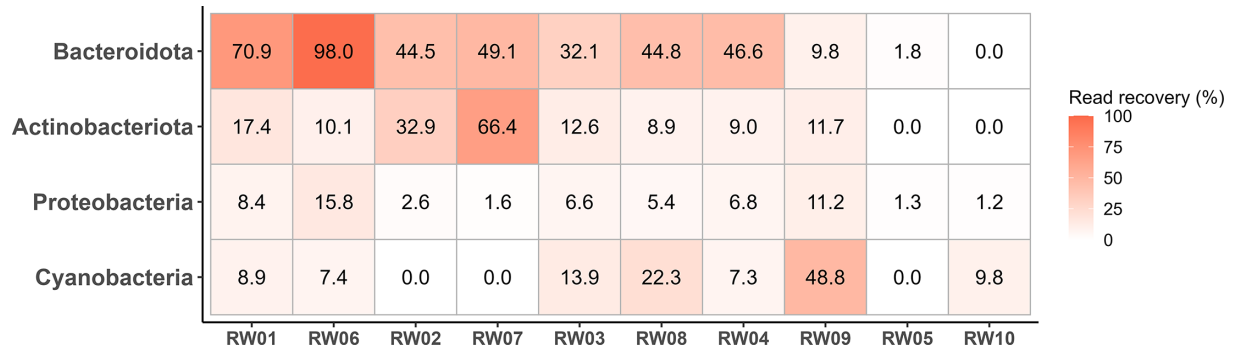


Fig. 3. Binning efficiencies within each sample for the four major phyla in this study. Samples are clustered according to sampling date. The numbers within heatmap cells represent binning efficiencies, which here is defined as the percentage of reads for each phylum which could be aligned to the MAG datasets.

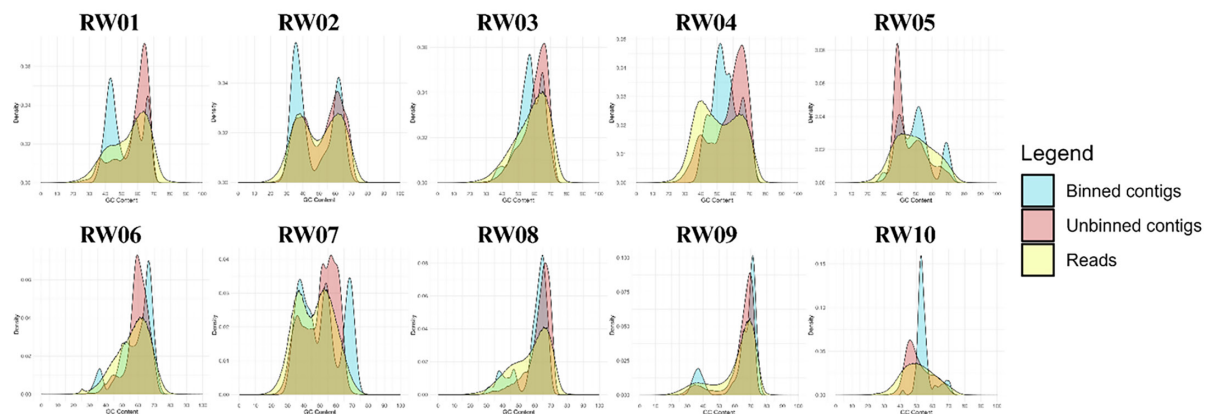


Fig. 4. Density plots showing contig and read GC content. The plot includes reads, binned contigs $\geq 1,000$ bp and unbinned contigs $\geq 1,000$ bp collected from all samples. Average GC content is displayed on the x-axis, and plot density is displayed on the y-axis.

The apparent overall higher binning efficiencies of the low GC *Bacteroidota* drove an investigation into the effect of GC content on assembly and binning. GC content was determined for binned and unbinned contigs and reads. Read GC content trended towards 60–65%. Density plots suggested a binning preference for reads with GC content between 55 and 75% (Fig. 4). This trend was consistent in most samples, except samples RW05 and RW10, obtained during summer. It should be noted that samples RW05 and RW10 experienced the lowest classification rates, assembly N50 values and binning efficiencies. The composition of the unclassified read fraction in these samples could not be determined from the database used; however, the convergence of these metrics suggests these samples may have harboured a substantial non-prokaryotic community, and patterns observed within them are interpreted with caution. No obvious trend was observed between GC content and binned contigs, though the density of binned contigs was generally higher than unbinned contigs for GC contents between 35–45%.

To investigate a preference during binning for specific taxa, reads with taxonomic annotations were aligned to MAG datasets. The proportion of aligned vs unaligned reads for each genus in each sample was calculated, as well as the average GC content of the reads for each genus (Fig. 5). A slight trend was observed which coincided with lower read GC content and greater MAG recovery. This trend was more pronounced in *Actinobacteriota* and *Bacteroidota*. In particular, *Actinobacteriota* saw poor MAG recovery as GC content trended towards 70%. Conversely, *Cyanobacteria* saw poor MAG recovery as GC content trended towards 40–45%.

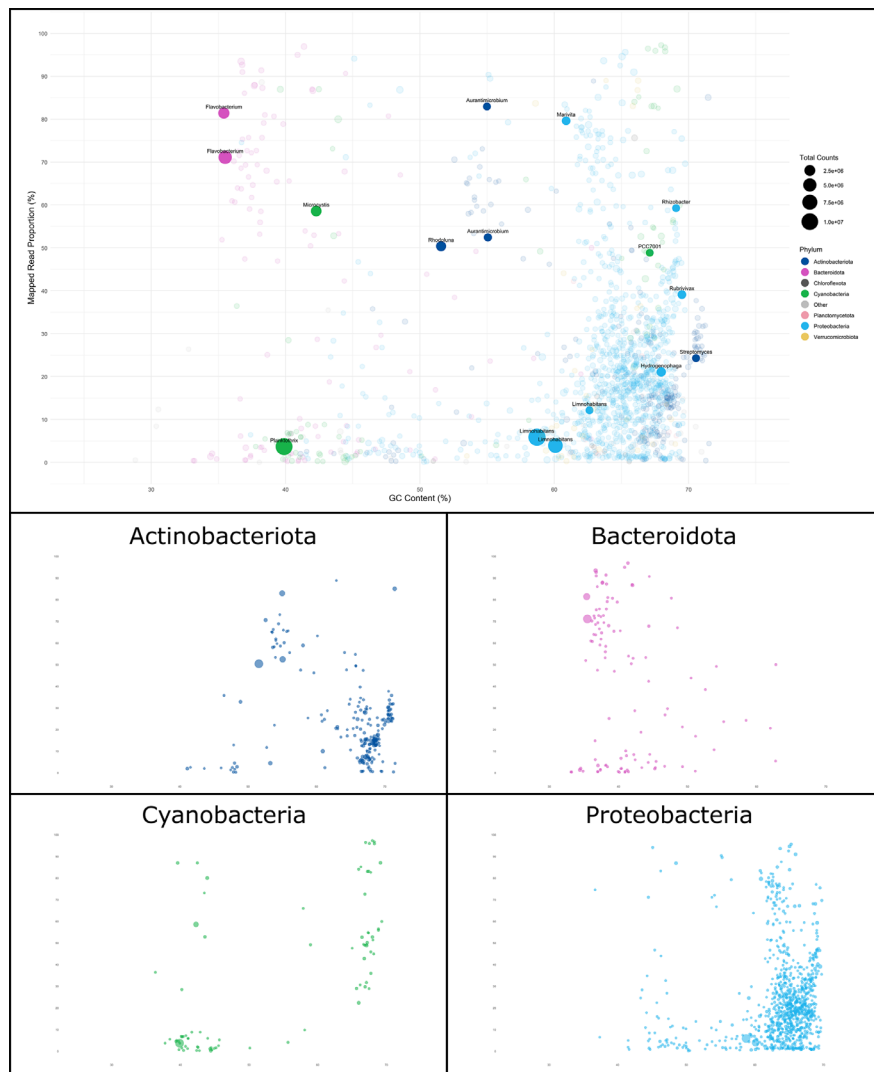


Fig. 5. Genus-specific binning efficiencies and GC content. Each bubble in the plot represents reads classified as one genus from one sample. Only genera with at least 10,000 reads are represented in the plot, and genera with at least 1,000,000 reads are highlighted and labelled. The y-axis represents the proportion of aligned vs unaligned reads, meaning higher bubble placement equates to greater MAG recovery. The x-axis represents average read GC content. Bubble size approximates to the sum of mapped and unmapped reads. Additional plots for the four major phyla (*Bacteroidota*, *Actinobacteriota*, *Proteobacteria* and *Cyanobacteria*) were drawn separately to highlight phyla-specific trends.

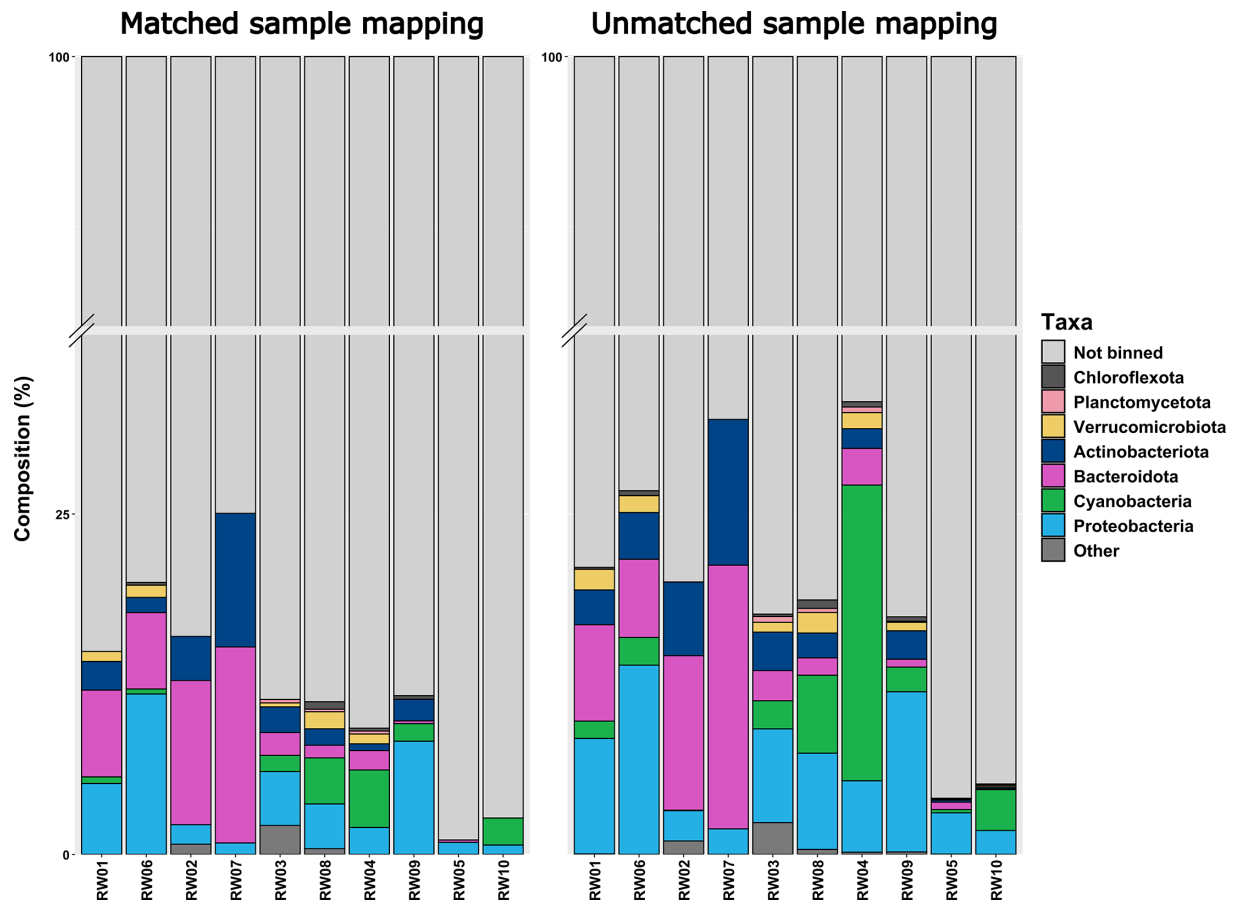


Fig. 6. Comparison of matched vs unmatched sample mapping. Composition (%) is on the y-axis and is calculated as the percentage of reads annotated with bacterial classifications. Samples are grouped according to sampling dates. Unmatched sample mapping saw an overall greater recovery in every sample, with the greatest recovery occurring in sample RW04, mostly attributed to *Cyanobacteria*.

The most substantial losses in taxa were for the genus *Planktothrix* from the phylum *Cyanobacteria* and the genus *Limnohabitans* from the phylum *Proteobacteria*. Notably, *Limnohabitans* saw substantial abundance (>1,000,000 reads) and poor MAG recovery in three different samples (RW02, RW03 and RW07). Conversely, *Flavobacterium*, the major genus from the *Bacteroidota* phylum, saw high MAG recoveries (>70%) in samples RW02 and RW07.

Taxonomic recovery through unmatched sample mapping

To recover taxonomic information lost through assembly and binning, reads from each sample were aligned to the total MAG dataset. On average, an additional 7.3% of classified bacterial reads per sample could be aligned to MAGs recovered from other samples, representing taxonomic information that would otherwise be absent from per-sample MAG datasets (Fig. 6 and Data S7). The greatest recovery occurred in sample RW04 (24%), primarily due to *Cyanobacteria* (17.5%). *Cyanobacteria* also had the greatest recovery summed across samples (25%), followed by *Proteobacteria* (23.9%), *Bacteroidota* (9.6%), *Actinobacteriota* (9.1%) and *Verrucomicrobiota* (3.1%).

Planktothrix showed the most substantial recovery due to unmatched sample mapping. The MAG recovered from sample RW09 had a significantly higher abundance in sample RW04 (0.35 and 12.26%, respectively), wherein it also exhibited its highest coverage (~460×). Additionally, 42 other MAGs exhibited greater relative abundance outside of the samples they were originally recovered.

DISCUSSION

As the recovery of MAGs becomes standard practice for studying microbial ecology, it is important to recognize how it may skew our understanding of microbial diversity. Here, we show that species abundance does not equate to the ease of MAG recovery, as several genera with major representation (namely *Planktothrix* and *Limnohabitans*) were absent from the medium and high-quality

MAG datasets. Additionally, we show that species from the *Bacteroidota* phylum were well-represented, as were low-GC species from the *Actinobacteriota*.

A key limitation of this study is the absence of independent, non-sequencing validation of community composition. Metagenomic read-level abundance estimates are subject to sequencing depth limitations and GC-associated bias and should not be interpreted as an absolute measure of true community composition. The biases identified here therefore represent a downstream, assembly- and binning-stage layer of distortion compounding an already imperfect read-level signal, which is a caveat we believe is important to acknowledge given the central role of read-based abundance in the study narrative.

The chosen assembly and binning method yielded 216 medium- and high-quality MAGs, which is an outcome comparable to other studies recovering MAGs from water samples [47, 48]. GC content of the recovered cyanobacterial MAGs showed high interquartile range, which is consistent with the literature suggesting fitness within this phylum for a range of ecophysiological conditions [49, 50]. Although studies suggest low-GC micro-organisms recovered from metagenomic data experience lower rates of genome completion [15], this was not immediately apparent in this study, and indeed lower GC organisms in the *Bacteroidota* and *Actinobacteriota* phyla experienced greater MAG recovery.

Although <25% of bacterial reads were recovered into MAGs in any sample, this loss is expected and consistent with other studies [51, 52]. Samples RW05 and RW10 exhibited the lowest bacterial classification rates, assembly N50 values and binning efficiencies, while also being the only samples whose GC content trended below 50%. Read level taxonomic profiling indicated a high proportion of unclassified reads in these samples. While the precise composition of this fraction could not be determined from the database used, the convergence of low bacterial classification, poor assembly metrics and sub-50% GC content suggests that these samples may have harboured a substantial non-prokaryotic community. Patterns observed within these samples are therefore interpreted with caution and are not considered representative of GC-associated MAG recovery bias in prokaryotes. Where possible, future studies could consider removal of eukaryotic organisms prior to DNA extraction to increase the robustness of MAG recovery efforts [53].

A major finding in this study was that highly abundant micro-organisms in the *Planktothrix* and *Limnohabitans* genera were not recovered into MAGs. This finding contrasts with the notion that MAG recovery efforts are biased towards more abundant organisms [54, 55]. Failure to recover a *Planktothrix* MAG is likely due to a poor assembly, evidenced by a higher density for unassembled reads around the GC content expected for *Planktothrix* (~40%) (Fig. 4; RW04). Additionally, unpublished work attempting to force *Planktothrix* MAG recovery was performed, wherein *Planktothrix*-classified reads from sample RW04 were extracted and assembled. The resulting assembly had a total length of 5.6 Mb; however, only 3.4 Mb was contained in contigs $\geq 1,000$ bp, the minimum size requirement for binning. Given the *Planktothrix* genome size range is between 4.7 and 6.7 Mb [56, 57], the assembly was considered inadequate for binning purposes. Nevertheless, a *Planktothrix* MAG was recovered, but it had a completion score <50% and was therefore discarded. The domination of an AT-rich organism such as *Planktothrix* is known to be problematic during sequencing [15] and is likely the reason a *Planktothrix* MAG was not recovered in this sample.

Despite the substantial *Limnohabitans* presence in three samples (Fig. 5), a representative MAG was not recovered. *Limnohabitans* is a diverse genus, and reads were attributed to 30 different *Limnohabitans* species (data not shown). Inspection of metabat2 bins in samples RW02 and RW07 revealed three *Limnohabitans* bins with completeness and contamination scores outside of the acceptable range for MAGs, including one bin with a contamination score of 383% (Data S7). Despite observably different morphologies and phenotypes, *Limnohabitans* species exhibit low genetic diversity [58], making it particularly susceptible to chimeric binning as contigs from divergent strains share sufficient sequence similarity to be incorrectly co-binned.

Notably, *Limnohabitans* species can indicate a collapsed cyanobacterial or algal bloom, as the genus thrives on the exudates of phototrophic microbes [59]. These ecological disturbances, as well as others such as eutrophication or sediment mixing, create niche conditions which drive strain-level genetic diversity [60–62]. Results collected here suggest microbes with pertinent functional roles may be lost due to MAG heterogeneity in similar studies, especially those with higher GC content. GC content is known to bias sequencing efforts, as GC-rich and GC-poor sequences respond poorly to library preparation and amplification protocols required by next-generation sequencing platforms [10, 63].

A trend was observed between lower GC content and higher MAG recovery in *Bacteroidota* and *Actinobacteriota*. Notably, the genus *Flavobacterium* in the phylum *Bacteroidota* demonstrated a high recovery rate in two samples (RW02; 83% and RW07; 73%). The MAGs recovered from this genus exhibited low GC ranging from 32 to 37%, while also exhibiting small genome sizes, averaging 2.1 Mb. Similarly, the efficiently recovered *Aurantimicrobium* and *Rhodulana* genera from the phylum *Actinobacteriota* exhibited a combined average GC content of 54% and genome size of 1.2 Mb, compared to 62% and 2.2 Mb in other *Actinobacteriota* genera, which were poorly recovered. These findings suggest that while GC content does not directly impact unbiased binning algorithms, it does influence genome size, facilitating the recovery of MAGs with lower GC content. This is consistent with recent findings linking lower GC content to smaller chromosome sizes in bacteria [64]. Collectively, these results indicate that MAG recovery efforts may be biased towards smaller, AT-rich genomes, which in this study resulted in the over-representation of species in the *Bacteroidota* and *Actinobacteriota*. Future studies could explore the influence of GC

content and genome size across more diverse real-world samples to understand the broader implications of this bias during MAG recovery.

We demonstrate that taxa can be recovered by mapping unmatched samples against the entire MAG dataset. By measuring the relative abundance of MAGs in this manner, we identified *Planktothrix* as the causative agent of a late summer bloom at the Inlet site. This identification is consistent with independent 16S rRNA amplicon sequencing of the same environment and overlapping sampling period, which reported *Planktothrix* as a dominant taxon during late summer [18], providing independent biological support for the read-level abundance profiles central to this study's narrative. While co-assembly is another method to resolve genomes present across multiple samples, it requires substantial RAM [9] which may not be feasible in all circumstances. Additionally, co-assemblies have the potential to produce more heterogeneous or erroneous MAGs by binning similar genomic fragments from distinct organisms [65]. The approach demonstrated here will enable other researchers facing similar challenges to resolve more taxa, despite restricted computational infrastructure.

The recovery of MAGs from metagenomic sequencing data has become standard practice in microbial ecology, significantly enhancing our understanding of microbial diversity. In this study, we demonstrate that species abundance does not necessarily correlate with ease of MAG recovery. Our findings show the efficient recovery of smaller, low-GC microbes, particularly over-representing species from the *Bacteroidota* and *Actinobacteriota*. Such identifications are crucial for addressing questions related to functions such as substrate use and microbial interactions. While MAG recovery is a powerful and essential technique, we hope this study encourages microbial ecologists to adopt similar investigations during MAG recovery efforts. Such practices will ultimately provide insights into how MAG recovery may skew the representation of specific taxa on a global scale. It should be noted that the mean number of reads obtained in this study (42.9 M) was lower than that obtained in similar studies comparing different binning approaches. Increased sequencing depth undoubtedly improves assembly and MAG recovery efforts and should be considered if the pipeline in this study is employed.

In conclusion, this study demonstrates that species abundance does not equate to ease of MAG recovery, with highly abundant genera such as *Planktothrix* and *Limnolobus* absent from the MAG dataset despite their dominance in read-based community profiles. Recovery bias was observed towards smaller, lower-GC genomes, particularly those of *Bacteroidota* and *Actinobacteriota*, and is likely driven by a combination of GC-associated sequencing bias, genome size and strain-level heterogeneity acting as compounding determinants of assembly and binning success. We further demonstrate that unmatched sample mapping can partially recover taxonomic information lost during per-sample MAG recovery. The pipeline is deployable on modest workstation-based infrastructure, making it accessible to research groups without access to high-performance computing facilities. We hope this study and the accompanying pipeline encourage microbial ecologists to critically evaluate MAG recovery outputs and to recognize that the organisms most ecologically significant in a given system may not always be those most amenable to genome-resolved analysis.

Funding information

The authors received no specific grant from any funding agency.

Acknowledgements

Funding for this study was gained through the NSW Environmental Trust (grant reference number 2018/RD/0068), the Australian Research Council Centre of Excellence in Synthetic Biology (CE200100029) and the Australian Research Council's Linkage Projects funding scheme (LP180100835). D.J.N. found financial assistance through the University of Newcastle Vice Chancellor's Higher Degree by Research Training Program.

Author contributions

D.J.N. and B.A.N. identified the knowledge gap and conceptualized the work. D.J.N. and C.R. processed the samples and generated the data. D.J.N. performed bioinformatics and data analysis and generated images. D.J.N. drafted the manuscript. B.A.N. and V.J.T. supervised the project. T.N. and V.J.T. reviewed and edited the manuscript. All authors helped with feedback and revision and accepted the final manuscript.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

The authors declare no personal or financial conflicts of interest which could have appeared to influence the work or writing of this manuscript.

References

1. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, et al. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci Rep* 2021;11:3030.
2. Nam NN, Do HDK, Loan Trinh KT, Lee NY. Metagenomics: an effective approach for exploring microbial diversity and functions. *Foods* 2023;12:2140.
3. Goussarov G, Mysara M, Vandamme P, Van Houdt R. Introduction to the principles and methods underlying the recovery of metagenome-assembled genomes from metagenomic data. *Microbiologyopen* 2022;11:e1298.
4. Nayfach S, Roux S, Seshadri R, Udworthy D, Varghese N, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2021;39:499–509.
5. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2017;2:1533–1542.

6. Viver T, Conrad RE, Orellana LH, Urdiain M, González-Pastor JE, et al. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. *ISME J* 2021;15:1178–1191.
7. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020;36:1925–1927.
8. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J* 2021;19:6301–6314.
9. Zhou Y, Liu M, Yang J. Recovering metagenome-assembled genomes from shotgun metagenomic sequencing data: methods, applications, challenges, and opportunities. *Microbiol Res* 2022;260:127023.
10. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation sequencing data on *de novo* genome assembly. *PLoS One* 2013;8:e62856.
11. Khuu P, Sandor M, DeYoung J, Ho PS. Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proc Natl Acad Sci USA* 2007;104:16528–16533.
12. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. *Genomics* 2016;107:1–8.
13. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 2012;40:e72–e72.
14. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. Characterizing and measuring bias in sequence data. *Genome Biol* 2013;14:R51.
15. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, et al. GC bias affects genomic and metagenomic reconstructions, under-representing GC-poor organisms. *Gigascience* 2020;9:giaa008.
16. Liew KJ, Shahar S, Shamsir MS, Shaharuddin NB, Liang CH, et al. Integrating multi-platform assembly to recover MAGs from hot spring biofilms: insights into microbial diversity, biofilm formation, and carbohydrate degradation. *Environ Microbiome* 2024;19:29.
17. Goldstein S, Beka L, Graf J, Klassen JL. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genom* 2019;20:23.
18. Foysal MJ, Timms V, Neilan BA. Dynamics of the benthic and planktic microbiomes in a *Planktothrix*-dominated toxic cyanobacterial bloom in Australia. *Water Res* 2024;249:120980.
19. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
20. Bragg L, Tyson GW. Metagenomics using next-generation sequencing. *Methods Mol Biol* 2014;1096:183–201.
21. Borderes M, Gasc C, Prestat E, Galvão Ferrarini M, Vinga S, et al. A comprehensive evaluation of binning methods to recover human gut microbial species from a non-redundant reference gene catalog. *NAR Genom Bioinform* 2021;3:lqab009.
22. Jégousse C, Vannier P, Groben R, Glöckner FO, Marteinsson V. A total of 219 metagenome-assembled genomes of microorganisms from Icelandic marine waters. *PeerJ* 2021;9:e11112.
23. Vosloo S, Huo L, Anderson CL, Dai Z, Sevilano M, et al. Evaluating *de novo* assembly and binning strategies for time-series drinking water metagenomes. *Microbiol Spectr* 2021;9:e0143421.
24. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016;32:1088–1090.
25. Kang DD, Li F, Kirton E, Thomas A, Egan R, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;7:e7359.
26. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;32:605–607.
27. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;25:1043–1055.
28. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;35:725–731.
29. Roux S, Trubl G, Goudeau D, Nath N, Couradeau E, et al. Optimizing *de novo* genome assembly from PCR-amplified metagenomes. *PeerJ* 2019;7:e6902.
30. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 2019;35:4162–4164.
31. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
32. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.
33. Edgar RC. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* 2022;13:6968.
34. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.
35. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
36. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021;49:W293–W296.
37. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019;20:257.
38. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;50:D785–D794.
39. Breitwieser FP, Salzberg SL. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* 2020;36:1303–1304.
40. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 2013.
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
42. Wilkinson L. ggplot2: elegant graphics for data analysis by WICKHAM. H: Oxford University Press. 2011.
43. Kolde R, Kolde MR. Package 'pheatmap'. R package 1; 2015
44. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;27:824–834.
45. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–1676.
46. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018;6:158.
47. Dalcin Martins P, Frank J, Mitchell H, Markillie LM, Wilkins MJ. Wetland sediments host diverse microbial taxa capable of cycling alcohols. *Appl Environ Microbiol* 2019;85:e00189-19.
48. Xue Z, Han Y, Tian W, Zhang W. Metagenome sequencing and 103 microbial genomes from ballast water and sediments. *Sci Data* 2023;10:536.
49. Hu E-Z, Lan X-R, Liu Z-L, Gao J, Niu D-K. A positive correlation between GC content and growth temperature in prokaryotes. *BMC Genomics* 2022;23:110.

50. Prabha R, Singh DP, Sinha S, Ahmad K, Rai A. Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. *Mar Genomics* 2017;32:31–39.
51. Diamond S, Andeer PF, Li Z, Crits-Christoph A, Burstein D, et al. Mediterranean grassland soil C-N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nat Microbiol* 2019;4:1356–1367.
52. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, et al. Genome-centric view of carbon processing in thawing permafrost. *Nature* 2018;560:49–54.
53. Demkina A, Slonova D, Mamontov V, Konovalova O, Yurikova D, et al. Benchmarking DNA isolation methods for marine metagenomics. *Sci Rep* 2023;13:22138.
54. Abdulkadir N, Saraiva JP, Schattenberg F, Toscan RB, Borim Correa F, et al. Combining flow cytometry and metagenomics improves recovery of metagenome-assembled genomes in a cell culture from activated sludge. *Microorganisms* 2023;11:175.
55. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013;31:533–538.
56. McKindles KM, McKay RM, Bullerjahn GS. Genomic comparison of *Planktothrix agardhii* isolates from a Lake Erie embayment. *PLoS One* 2022;17:e0273454.
57. Pancrace C, Barny M-A, Ueoka R, Calteau A, Scalvenzi T, et al. Insights into the *Planktothrix* genus: genomic and metabolic comparison of benthic and planktic strains. *Sci Rep* 2017;7:41181.
58. Kasalický V, Jezbera J, Hahn MW, Šimek K. The diversity of the *Limnohabitans* genus, an important group of freshwater bacterioplankton, by characterization of 35 isolated strains. *PLoS One* 2013;8:e58209.
59. Jezberová J, Jezbera J, Znachor P, Nedoma J, Kasalický V, et al. The *Limnohabitans* genus harbors generalistic and opportunistic subtypes: evidence from spatiotemporal succession in a canyon-shaped reservoir. *Appl Environ Microbiol* 2017;83.
60. Galand PE, Lucas S, Fagervold SK, Peru E, Pruski AM, et al. Disturbance increases microbial community diversity and production in marine sediment. *Front Microbiol* 2016;7:1950.
61. Goyal A, Bittleston LS, Leventhal GE, Lu L, Cordero OX. Interactions between strains govern the eco-evolutionary dynamics of microbial communities. *eLife* 2022;11.
62. Herold M, Martínez Arbas S, Narayanasamy S, Sheik AR, Kleine-Borgmann LAK, et al. Integration of time-series metagenomics data reveals how microbial ecosystems respond to disturbance. *Nat Commun* 2020;11:5281.
63. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods* 2009;6:291–295.
64. Alimpanis A, Swain M, Gatherer D, McEwan N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb Genom* 2018;4:e000168.
65. Shaiber A, Eren AM. Composite metagenome-assembled genomes reduce the quality of public genome repositories. *mBio* 2019;10.

The Microbiology Society is a membership charity and not-for-profit publisher.

Your submissions to our titles support the community – ensuring that we continue to provide events, grants and professional development for microbiologists at all career stages.

Find out more and submit your article at microbiologyresearch.org