

<b>Manuscript Number:</b>	GIGA-D-23-00067R2	
<b>Full Title:</b>	Metaphor - A workflow for streamlined assembly and binning of metagenomes	
<b>Article Type:</b>	Technical Note	
<b>Funding Information:</b>	Australian Research Council (DE220100965)	Dr Vanessa Rossetto Marcelino
	National Health and Medical Research Council (GNT1159458)	Prof. Kim-Anh Le Cao
	Australian Research Council (DP200101613)	A/Prof Heroen Verbruggen
<b>Abstract:</b>	<p>Recent advances in bioinformatics and high-throughput sequencing have enabled the large-scale recovery of genomes from metagenomes. This has the potential to bring important insights as researchers can bypass cultivation and analyse genomes sourced directly from environmental samples. There are, however, technical challenges associated with this process, most notably the complexity of computational workflows required to process metagenomic data, which include dozens of bioinformatics software tools, each with their own set of customisable parameters that affect the final output of the workflow. At the core of these workflows are the processes of assembly - combining the short input reads into longer, contiguous fragments (contigs), and binning - clustering these contigs into individual genome bins. The limitations of assembly and binning algorithms also pose different challenges depending on the selected strategy to execute them. Both of these processes can be done for each sample separately or by pooling together multiple samples to leverage information from a combination of samples. Here we present Metaphor, a fully-automated workflow for genome-resolved metagenomics (GRM). Metaphor differs from existing GRM workflows by offering flexible approaches for the assembly and binning of the input data, and by combining multiple binning algorithms with a bin refinement step to achieve high quality genome bins. Moreover, Metaphor generates reports to evaluate the performance of the workflow. We showcase the functionality of Metaphor on different synthetic datasets, and the impact of available assembly and binning strategies on the final results.</p>	
<b>Corresponding Author:</b>	Kim-Anh Le Cao The University of Melbourne Parkville, VIC AUSTRALIA	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>	The University of Melbourne	
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Vinicius W. Salazar	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Vinicius W. Salazar	
	Babak Shaban	
	Maria del Mar Quiroga	
	Robert Turnbull	
	Edoardo Tescari	
	Vanessa Rossetto Marcelino	
	Heroen Verbruggen	

	Kim-Anh Le Cao
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	Minor editing amendments have been addressed. All latex files have been uploaded
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using</p>	Yes

a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)?



TECHNICAL NOTE

# Metaphor - A workflow for streamlined assembly and binning of metagenomes

Vinícius W. Salazar<sup>1</sup>, Babak Shaban<sup>2,†</sup>, Maria del Mar Quiroga<sup>2</sup>, Robert Turnbull<sup>2</sup>, Edoardo Tescari<sup>2</sup>, Vanessa Rossetto Marcelino<sup>3,4,5,6</sup>, Heroen Verbruggen<sup>5,§</sup> and Kim-Anh Lê Cao<sup>1,§,\*</sup>

<sup>1</sup>Melbourne Integrative Genomics, School of Mathematics & Statistics, University of Melbourne, Parkville, Victoria, Australia and <sup>2</sup>Melbourne Data Analytics Platform (MDAP), University of Melbourne, Parkville, Victoria, Australia and <sup>3</sup>Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria, Australia and <sup>4</sup>Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, Australia and <sup>5</sup>School of BioSciences, University of Melbourne, Parkville, Victoria, Australia and <sup>6</sup>Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Parkville, Victoria, Australia.

\*kimanh.lecao@unimelb.edu.au

<sup>†</sup>Deceased.

<sup>§</sup>Contributed equally.

## Abstract

Recent advances in bioinformatics and high-throughput sequencing have enabled the large-scale recovery of genomes from metagenomes. This has the potential to bring important insights as researchers can bypass cultivation and analyse genomes sourced directly from environmental samples. There are, however, technical challenges associated with this process, most notably the complexity of computational workflows required to process metagenomic data, which include dozens of bioinformatics software tools, each with their own set of customisable parameters that affect the final output of the workflow. At the core of these workflows are the processes of assembly - combining the short input reads into longer, contiguous fragments (contigs), and binning - clustering these contigs into individual genome bins. The limitations of assembly and binning algorithms also pose different challenges depending on the selected strategy to execute them. Both of these processes can be done for each sample separately or by pooling together multiple samples to leverage information from a combination of samples. Here we present Metaphor, a fully-automated workflow for genome-resolved metagenomics (GRM). Metaphor differs from existing GRM workflows by offering flexible approaches for the assembly and binning of the input data, and by combining multiple binning algorithms with a bin refinement step to achieve high quality genome bins. Moreover, Metaphor generates reports to evaluate the performance of the workflow. We showcase the functionality of Metaphor on different synthetic datasets, and the impact of available assembly and binning strategies on the final results.

**Key words:** Bioinformatics; pipeline; MAGs; Snakemake; high-throughput sequencing; microbial genomics

## Introduction

Genome-resolved metagenomics (GRM) is a set of techniques for the recovery of genomes from high-throughput sequencing data.

Applications of GRM have led to unprecedented insight into microbial diversity, ecology, and evolution, due to the recovery of (mostly uncultivated) metagenome-assembled genomes (MAGs)

Compiled on: June 19, 2023.

Draft manuscript prepared by the author.



[1, 2, 3, 4]. MAGs are essentially “bins” of contigs that are clustered together based on differential coverage and sequence composition; a bin is considered a MAG when it displays a high degree of completeness and a low degree of redundancy/contamination, which is usually calculated through the presence of marker genes in the bin. Advances in GRM have consistently improved the quality of recovered MAGs, and large-scale studies reconstructing and analysing thousands of MAGs have become prominent in microbiology research. Even with the inherent biases that accompany the generation of MAGs, it is evident that the benefits outweigh the risks, and researchers are increasingly in need of automated data processing methods for assembling and binning metagenomes [5]. Data pipelines that perform such experiments are inherently complex, have high computing cost, use heterogeneous data sources, have dozens of customisable parameters, and depend on several specialised bioinformatics software [6, 7].

An additional domain-specific challenge for GRM studies is the strategy used for assembling and binning each sequenced sample. Data (raw reads generated by the sequencer) originating from multiple samples may be assembled separately or pooled together, depending whether they come from the same population, specimen, or environment. This results in either a set of contigs for each sample or a ‘coassembly’ of the pooled samples. Similarly, in the metagenome binning step, where contigs are clustered into genome bins, one may do this individually for each set of assembled contigs, or by pooling together contigs from multiple samples and then mapping each individual sample to this catalogue of contigs (‘cobinning’) [8]. The latter approach allows binning algorithms to account for differential coverage of contigs across samples, enriching the information available for clustering. The chosen strategy for assembly and binning may have important consequences for the final results, *i.e.*, the quality of the assembly and of the recovered bins [8]. It is hypothesised that pooled assembly and binning may lead to improved results when analysing communities with high genetic diversity, and to poorer results when there is a high level of intraspecies/strain-level diversity [9],

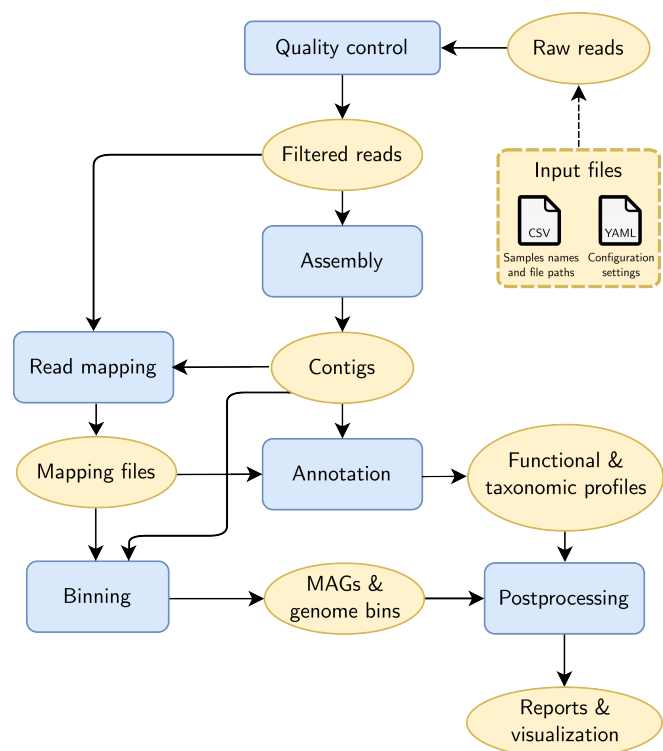
Here we present Metaphor, an automated and flexible workflow for the assembly and binning of metagenomes, which recovers prokaryotic genomes from metagenomes efficiently and with high sensitivity, and provides taxonomic and functional abundance data for quantitative metagenome analyses. Our software advances existing metagenomic pipelines by combining two core features: the usage of multiple binning software along with a binning refinement step, and the possibility of defining groups for assembly and binning of samples. This effectively allows scaling Metaphor to process multiple datasets in a single execution, performing assembly and binning in separate batches for each dataset, and avoiding the need for repeated executions with different input datasets. The workflow includes native functionality for downstream integration with ‘omics statistical toolkits [10, 11], so that abundance data can be easily imported into these tools, and with the Anvi’o [12] platform, which allows importing the collections of bins generated by Metaphor along with contig coverage data. Metaphor generates detailed performance metrics at the end of each module of the workflow to provide users with a high-level summary of their analysis, and has been designed to be user-friendly, portable, and flexible, as users can choose between different strategies for assembly and binning. We demonstrate its functionality using different synthetic datasets and discuss how these different strategies can impact data analyses in terms of quality of the resulting assembly and genome bins.

## Design and Implementation

Metaphor stands out from existing GRM pipelines by offering flexible options for assembly and binning combined with multiple binning software and a binning refinement step. See Ta-

ble 1 for a comparison of Metaphor’s features with other state-of-the-art GRM workflows. The workflow is implemented with Snakemake [13], a widely-used scientific workflow management system. In each module, computing steps (called “rules” by Snakemake) consist of both third-party bioinformatics software [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28] and custom scripts that connect different parts of the workflow, listed on Table 2.

The workflow consists of six modules: quality control (QC), assembly, annotation, mapping, binning, and postprocessing. In the QC module, raw sequencing reads are filtered and trimmed. Metagenomic assembly is then performed. Coding sequences are predicted from the assembled contigs and used for functional and taxonomic annotation. The quality-filtered reads are mapped against the contigs, generating coverage statistics employed by the binning algorithms. After binning is complete, bins are refined and dereplicated. Lastly, the postprocessing module renders runtime and memory usage metrics and generates an HTML report. A simplified version of the flow of data between the different modules of the workflow is shown on Fig 1.



**Figure 1. Simplified workflow diagram.** Workflow modules are represented by rectangular blue shapes and data files are represented by oval yellow shapes, except for entrypoint files shown in a dashed yellow rectangle. Arrows indicate input and output of data between modules.

The choice of bioinformatics tools was informed by the results of the 2nd Critical Assessment for Metagenome Interpretation (CAMI II) [8, 36], striving for the maximum trade-off between performance, efficiency, and software sustainability. Although the latter is a subjective factor, selecting and streamlining dependencies with regard to code quality, maintenance, and community support is a critical factor when maintaining complex bioinformatics pipelines [6, 37]. Each third-party software (along with its version) is defined in an individual requirements file that is used by Snakemake to create a virtual environment and run that particular step. To facilitate citing these tools, Metaphor packages a `bibs/` directory containing all citations in the Bibtext format.

The workflow takes two files as input: a tab-delimited file con-

**Table 1.** Comparison of features between Metaphor and state-of-the-art GRM workflows as listed by [29]. Data adapted to include Metaphor.

Features	Metaphor v1.7.7	ATLAS [30]	MetaWRAP [31]	nf-core/mag [32]	MAGNETO [29]
Preprocessing					
Reads trimming	✓	✓	✓	✓	✓
Contamination	✓	✓	✓	✓	✓
Assembly					
Coassembly possible	✓		✓	✓	✓
Coassembly by groups	✓				
Compute sets to coassemble				✓	
Assembly evaluation	✓				
Binning					
Cobinning possible	✓		✓	✓	✓
Multiple binning software	✓	✓	✓		
Bin refinement	✓	✓	✓		
Bin reassembly		✓	✓		
Postprocessing					
MAGs quality check	✓	✓	✓	✓	✓
Dereplication step	✓	✓	✓	✓	✓
Genome annotation	✓	✓	✓	✓	✓
Gene catalogue	✓			✓	✓
HTML Report	✓	✓		✓	✓
Reproducibility					
Workflow management	✓	✓		✓	✓
Packages Management	✓	✓		✓	✓

**Table 2.** Modules, steps and software used in Metaphor.

Module	Step	Software
Quality Control (QC)	Trim adapters and filter low quality reads	fastp [14]
	Generate QC reports	FastQC [15]
	Combine QC reports	MultiQC [16]
Assembly	Assemble filtered and merged reads into contigs	MegaHit [17]
	Perform assembly evaluation	MetaQUAST [18]
	Assembly report and plots	Metaphor script*
Mapping	Map reads	MiniMap2 [19]
	Sort and index mapped reads	Samtools [20]
Annotation	Prediction of coding sequences from contigs	Prodigal [21]
	Annotation of coding sequences	Diamond, NCBI COG [22, 23]
	Annotation of MAGs	Prokka [24]
	Annotation report and plots	Metaphor script*
Binning	Cluster contigs into bins	VAMB [25]
	Cluster contigs into bins	MetaBAT2 [26]
	Cluster contigs into bins	CONCOCT [27]
	Dereplicate and score bins	DAS Tool [28]
	Binning report and plots	Metaphor script*
Postprocessing	Concatenate benchmarks	Metaphor script*
	Plot benchmarks	Metaphor script*

\* External libraries used in Metaphor scripts: [33, 34, 35].

**Table 3.** Datasets from CAMI II used to assess the workflow. Columns show the number of samples and size in gigabytes of each dataset, along with the amount of reference genomes used to generate the dataset

Dataset	Identifier	No. of samples	Size (GB)	No. reference genomes
Marine	marmg	10	50	622
Strain Madness	strmg	100	200	408
Human Airways	h_airways	10	44	1394
Human Genital	h_urogenital	9	39	1394
Human Gut	h_gastrointestinal	10	44	1057
Human Oral	h_oral	10	43	1057
Human Skin	h_skin	10	44	1394

**Table 4.** Output files for each strategy. If only one dataset/group is being analysed, assembly and binning results are named as “Coassembly” and “Cobinning” respectively. If multiple datasets/groups are used, the results are named according to the group/dataset’s name.

Strategy	Description	Reads files	Assemblies	Bins
SASB	Single assembly, Single binning	Sample_0.fastq	Sample_0_contigs.fasta	Sample_0_bins/
		Sample_1.fastq	Sample_1_contigs.fasta	Sample_1_bins/
		Sample_2.fastq	Sample_2_contigs.fasta	Sample_2_bins/
SACB	Single assembly, Cobinning	Sample_0.fastq	Sample_0_contigs.fasta	Cobinning_bins/
		Sample_1.fastq	Sample_1_contigs.fasta	
		Sample_2.fastq	Sample_2_contigs.fasta	
CACB	Coassembly, Cobinning	Sample_0.fastq	Coassembly_contigs.fasta	Cobinning_bins/
		Sample_1.fastq		
		Sample_2.fastq		

taining sample names and file paths to the raw reads, and a configuration file in the YAML format, which will set the workflow parameters (see Fig 1). These files can be automatically generated by Metaphor and edited by the user, or created from scratch. The output of Metaphor consists of a directory for each module, further subdivided into the rules within each module. This is described in detail in the documentation [38].

### Assessment on CAMI II synthetic datasets

To demonstrate the functionality of Metaphor, we analysed datasets from CAMI II [8]. All datasets consist of short and long reads generated by simulation of collections of reference genomes [39]. Only short reads were used for each dataset, as Metaphor does not yet support long reads. Specifically, we used the Marine metagenome dataset (identified as ‘marmg’), the Strain Madness dataset (identified as ‘strmg’), and the Human Microbiome dataset, which consists of five sets of samples, each corresponding to a different sampling location in the human body, which were treated as distinct datasets (3). The following strategies were employed for each dataset: single assembly, single binning (‘SASB’), where each sample is individually assembled and binned; single assembly, cobinning (‘SACB’), where each sample is assembled individually and then binned with other samples from the same dataset; coassembly, cobinning (‘CACB’), where all samples from the dataset were assembled and binned together. Table 4 illustrates how this works in practice, in terms of generated output files. Metaphor allows defining multiple groups for coassembly or cobinning to analyse multiple independent datasets with a single execution.

In order to assess the effect of different assembly strategies, we used MetaQUAST [18] to compare the assemblies generated by the workflow with the collections of reference genomes. For the different binning strategies, we compared metrics obtained from DAS Tool, the software used for dereplicating and evaluating genome

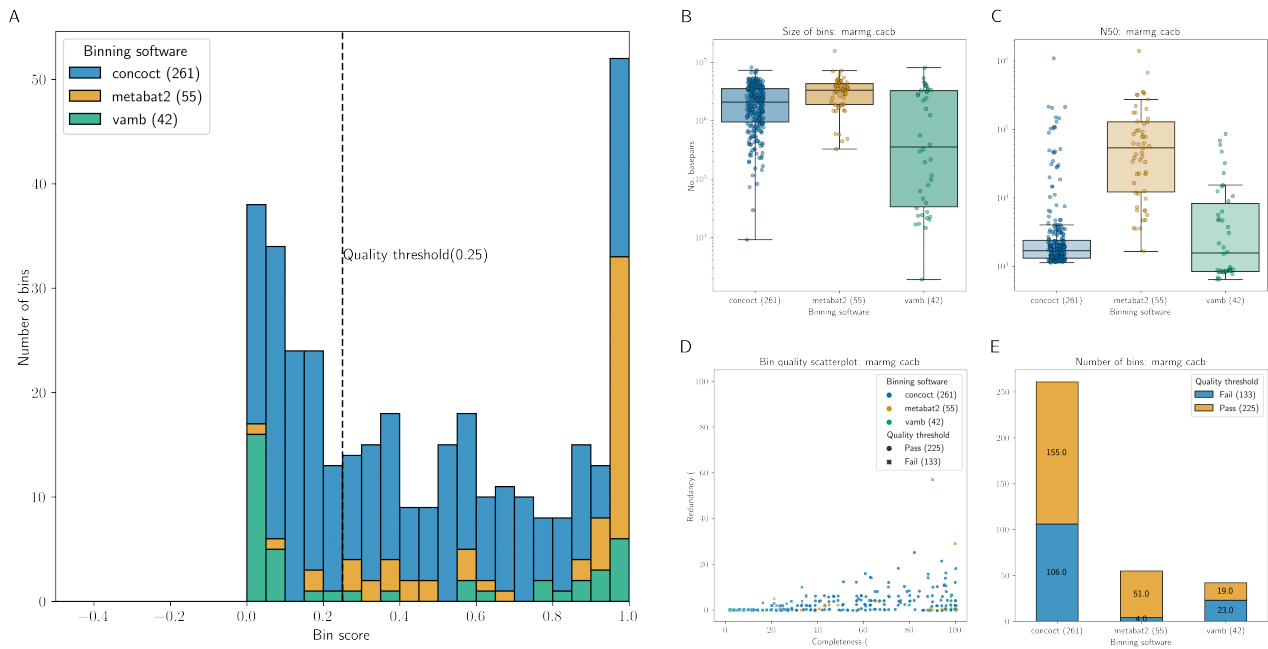
bins, after a second round of dereplication with dRep [40]. This is because data generated with the SASB strategy will likely result in redundant bins, as for that strategy there is no dereplication between samples and since samples within a dataset have similar composition, it is likely that a genome bin can be generated repeatedly by different samples. dRep performs dereplication based on the Average Nucleotide Identity between genomes, a metric which has been consistently used as a proxy to differentiate taxonomy at the species and strain levels [41]. dRep was run with default clustering parameters, and without any length, completeness, or contamination cutoffs. We used Spartan [42], the High Performance Computing (HPC) system at The University of Melbourne to run the pipeline. Jobs were dispatched to nodes with the SLURM scheduler, using up to 64 processors and 300 GB RAM per node.

## Results and Discussion

After running Metaphor on the CAMI II Marine, Strain Madness and Human Microbiome datasets, we illustrate the different outputs generated by the workflow, and compare the effects of different assembly and binning strategies on workflow performance.

### Reconstruction of metagenome-assembled genomes

Metaphor produces genome bins generated with three tools: Vamb, MetaBAT2 and CONCOCT [25, 26, 27] that are refined with DAS Tool [28]. DAS Tool performs bin refinement through a “dereplication, aggregation and scoring” process, in which candidate bins are initially scored based on the presence/absence of single-copy marker genes (SCGs, which are a proxy for bin completeness). Redundant candidate bin sets are then aggregated and an iterative scoring process is performed, so only the best-quality, non-redundant bins remain; the bin score ( $S_b$ ) increases with the number of SCGs and



**Figure 2.** Binning report generated by Metaphor for the CAMI II Marine metagenome dataset processed with the ‘CACB’ (coassembly, cobinning) setting. Panel A shows a stacked histogram of the distribution of bin scores, with the defined quality threshold highlighted as a dashed line. Panels B and C show, respectively, the size (in base pairs) and N50 of bins. The Y-axis is in log-scale. Panel D shows a scatterplot of completeness and redundancy for each bin. Colours indicate the tool used to generate the bin, and the symbols indicate whether that bin passed or failed the bin score quality threshold (corresponding to the same value in the dashed line of Panel A). Panel E shows how many bins passed or failed the quality threshold for each binning tool.

decreases with duplicate SCGs per bin. Please refer to [28], Figure 1 and Equation 1 for an overview of the DAS algorithm and the formula to determine the bin score, respectively. The input for each binning tool differs slightly, but they all rely on the catalogue of contigs obtained from the assembly and the coverage files obtained from the read mapping module (see Fig 1). A report is generated for each of the binning groups (only one is generated if cobinning is performed), which highlights three key metrics: completeness, redundancy, and bin score. The first two metrics are calculated by the presence/absence of single-copy genes, and the latter is a function of the former two. Plots generated by an example report are shown in Fig 2. It is possible to compare the performance of the different binning software and obtain the proportion of bins above a specified particular quality threshold based on the bin score. The source table for the report is provided, so that users can generate custom reports and inspect specific individual bins. Bins that pass the quality threshold are stored in individual FASTA files, so they can easily be used for downstream analyses with tools such as CheckM or GTDB-Tk [43, 44]. We chose not to include these software in the workflow as they rely on fairly large reference databases and/or contain several different steps that are dependent on third-party software, which would affect Metaphor’s portability. Bin collections generated with Metaphor can be imported into the Anvi’o along with coverage data (BAM files), so users can use the interactive interface of Anvi’o to examine the bins.

### Contig-level taxonomic and functional profiling

To facilitate quantitative metagenomics applications, Metaphor’s annotation module generates contig-level functional and taxonomic profiles based on the NCBI COG database [23]. These are obtained by predicting coding sequences with Prodigal and then aligning the resulting amino acid files with Diamond [21, 22] in the “iterative” mode. This setting performs repeated rounds of alignment, with an increasing degree of sensitivity when no hits are detected in the previous round. Abundances for each feature are

calculated based on the coverage of all coding sequences which align to that feature. Fig 3 illustrates the profile visualisations offered by Metaphor: a heatmap of COG categories for the functional profile and a stacked barplot for the most abundant taxa (for the latter, one plot is generated for each taxonomic rank). The annotation module outputs count tables with both absolute and relative abundance values of taxa and functional categories, and may be directly imported by downstream statistical toolkits such as MixOmics or PhyloSeq [10, 11].

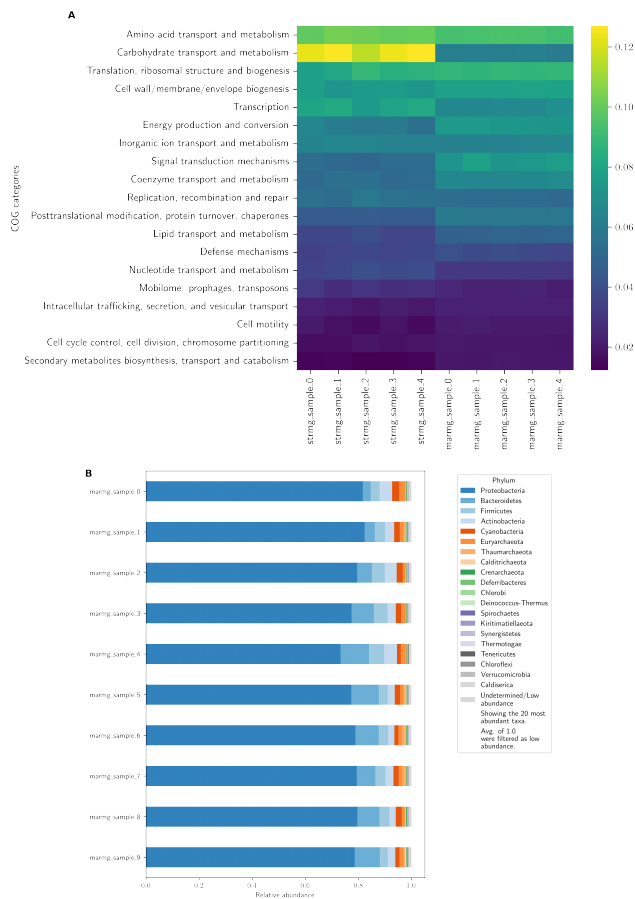
### Quality control and performance metrics

Additional outputs produced by Metaphor include the quality control reports from the fastp and FastQC tools, with a summary of FastQC outputs being produced by MultiQC [14, 15, 16]. A simple report is produced by the assembly module with sequence statistics of the assembled contigs (e.g. N50, number of contigs, total and mean length of contigs), and performance metrics. At the end of the workflow execution, the postprocessing module generates figures obtained from the “benchmark” files provided by Snake-make. These files contain process information such as runtime and memory consumption. Metaphor plots these metrics in two ways: total per rule and per-sample mean (Fig 4) as some rules run only once across all samples, while other rules run per sample. These plots help identify computational bottlenecks and assess whether computing resources are adequate.

### Assembly and binning strategies

The effects of distinct assembly and binning strategies on the final output of metagenomic workflows are highly dependent on the data source and research context [8]. As such, the choice of individual or group assembly and binning can only be assessed a posteriori. We compared three different strategies: single assembly and single binning (‘SASB’), single assembly and cobinning (‘SACB’), and coassembly and cobinning (‘CACB’), see Table 4 and Section ‘As-





**Figure 3. Annotation plots generated by Metaphor on the Strain Madness ('strmg') and the Marine ('marmg') datasets.** Panel A displays the functional profile as a heatmap of the relative abundance of functional COG categories (Y-axis) across samples (X-axis) for five samples from Strain Madness and Marine datasets. Panel B displays the taxonomic profile of the Marine dataset as a stacked barplot of relative abundance of taxa. In this case, the phylum rank was used, but Metaphor generates this for the most common taxonomic ranks (phylum, class, order, family, genus, species). The number of abundant taxa can be easily adjusted in the workflow settings. For both taxonomic and functional profiles, abundance of each feature is calculated from coverage values for each gene.

assessment on CAMI II synthetic datasets' for details. For assembly, we used the five different groups in the Human Microbiome dataset along with the Strain Madness and Marine datasets. We only used the latter two datasets for the binning assessment.

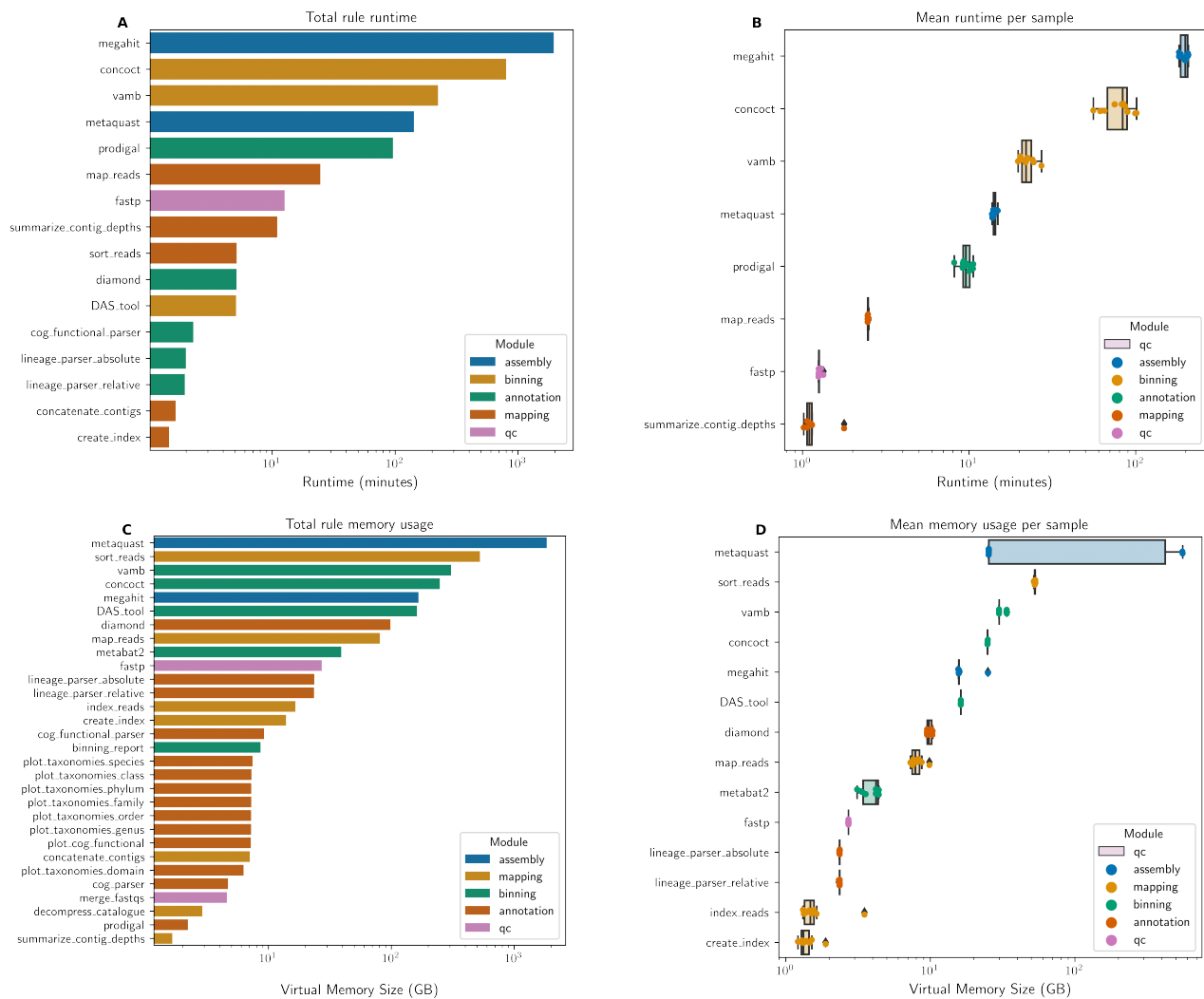
We used six metrics to evaluate assembly performance: percentage of recovered genome fraction, size of the largest contig, duplication ratio, length of misassembled contigs, number of misassemblies, and number of mismatches per 100 thousand base pairs. High values for the first two metrics and low values for the last four indicate better performance. We observed a general trade-off between assembly completeness (represented by the first two metrics), and the number of errors in the assembly (represented by the last four metrics) (Figure S1). In most datasets, assemblies were more complete and contiguous, albeit with more errors when the Coassembly strategy was used. The exception was the Strain Madness ('strmg') dataset, for which the Individual assembly was more complete and contiguous, albeit with more errors. This may be attributed to the high degree of strain/intraspecies diversity in that dataset [8]. A high degree of similarity between the related genomes likely confounds assembly algorithms, and pooling samples together may aggravate this effect [5].

To evaluate differences between binning strategies, we compared the number and quality of bins after refinement with DAS Tool. Bins generated with each approach were further dereplicated

with dRep [40]. This is because the SASB strategy generates a set of bins for each sample, and datasets with similar composition will likely generate redundant bins, as there is no dereplication of bins between samples. Results varied significantly between the Marine and Strain Madness datasets. In both datasets, the mean bin score was the highest for the CACB strategy (Figure S3). However, in the Strain Madness dataset, CACB produced a significantly lower number of bins (33 compared with 259 and 215 generated with SASB and SACB), which did not occur in the Marine dataset. The performance of each binning tool is also variable between strategies and is conditional on the characteristics of the original dataset, with no clear "winner", and each tool favouring particular performance metrics, in agreement with results from the 2nd CAMI Challenge [8]. Tools like DAS Tool attempt to conciliate the output of multiple binning algorithms to generate a consensus output which theoretically outperforms each individual algorithm.

Since the binning performance is assessed as a proxy of the combination of quantity and quality of generated bins, rather than only one metric or the other, we calculated the cumulative bin score (the sum of scores of all bins) and the number of bins above an increasing score threshold, shown on Fig 6. The higher the threshold, the more significant the differences between the cumulative scores, as only bins with the highest quality compose the score. For the Marine dataset, we observed a higher score and a larger number of bins in the CACB strategy and the exact opposite in the Strain Madness dataset. In both datasets, there was a clear difference between SASB before dereplication and the other strategies, confirming that several highly similar samples produce redundant bins. That difference was also present in the SACB strategy, albeit not so pronounced (see Figure S4 for the comparison of dereplicated and non-dereplicated data). This suggests that for both of these strategies, further dereplication is recommended [5]. Although the Strain Madness dataset shows fewer bins generated with CACB—a summary of the bins recovered with that dataset is displayed on ???. The cumulative bin score for that strategy remained similar to SACB and SASB above the 0.8 score threshold, since there are fewer bins with a score lower than that. In that same dataset, SASB showed the best performance, although differences were small above the 0.8 threshold. In the Marine dataset, there were more pronounced differences between strategies. CACB produced the larger quantity and higher cumulative score of bins, followed by SASB and SACB.

In summary, our results indicate that, for most metagenomic analysis scenarios, coassembly followed by cobinning is recommended, assuming that samples are sourced from a similar environment or population. The exception to this is when there is a high level of intraspecies/strain-level diversity across samples, like in the Strain Madness dataset. In that scenario, single assembly followed by single binning is preferred, followed by dereplication of bins between samples. There is, however, a trade-off between the different approaches, as computational requirements are higher for the pooled strategies. Coassembly resulted in higher genome recovery fractions and larger contigs, although usually at the expense of a higher number of misassemblies and higher duplication ratio. When combining coassembly with cobinning, there is a remarkable improvement in the quantity and quality of bins generated for a diverse dataset (represented by the Marine dataset), where the difference was negligible in the Strain Madness dataset. Therefore, when deciding the assembly and binning strategy, it is important to consider the expected strain-level diversity and abundances of each individual genome, as the interaction between these factors is likely to limit the resolution of recovered bins. This is shown in the CAMI II challenge [8] (see Figure 1g); genomes with low strain diversity (*i.e.* are less than 95% similar with any other genomes) have higher correlation between sequencing coverage and recovered fraction than common genomes ( $\geq 95\%$  similar to other genomes in the sample), although many times sequencing coverage was not all correlated with genome recovery fraction, specially for smaller bins that represent plasmids or circular elements.



**Figure 4. Performance metrics report generated by Metaphor on the Marine dataset processed with the SASB strategy.** Total runtime per rule (A), mean runtime per sample (B), total memory usage per rule (C), and mean memory usage per sample (D). X-axis is in log format. Cutoffs are applied to omit rules with short runtime or low memory usage. Colours indicate the workflow module of each rule.

## Availability and Future Directions

Metaphor is available through Bioconda [45], a popular repository of bioinformatics software. It can be installed with a single command from the conda package manager [46] or from source using pip, the Python package manager. The installation of all third-party software used by Metaphor is handled automatically by Snakemake and conda. It can be easily deployed in different computing environments, such as high performance computing clusters and cloud instances, due to Snakemake's support of execution profiles. Metaphor is developed with documented best practices in workflow development [6, 47], striving for reproducibility and transparency of its results. Data used for the testing Metaphor's installation (see documentation for details) is available from GitHub at <https://github.com/vinisalazar/mg-example-data>. This data is a subset of the CAMI I challenge data [36] that is reduced in size in order to run test commands in a reasonable time.

The workflow may be extended to support downstream tools such for genome analysis such as GTDB-Tk, CheckM, and dRep. This may help with further improvement of strain-level resolution in bins; there are a number of strategies for that, such as identification of misassembled contigs or using the assembly graph for variant detection [48, 49]. New functionality may also be added for the identification of eukaryotic and viral contigs;

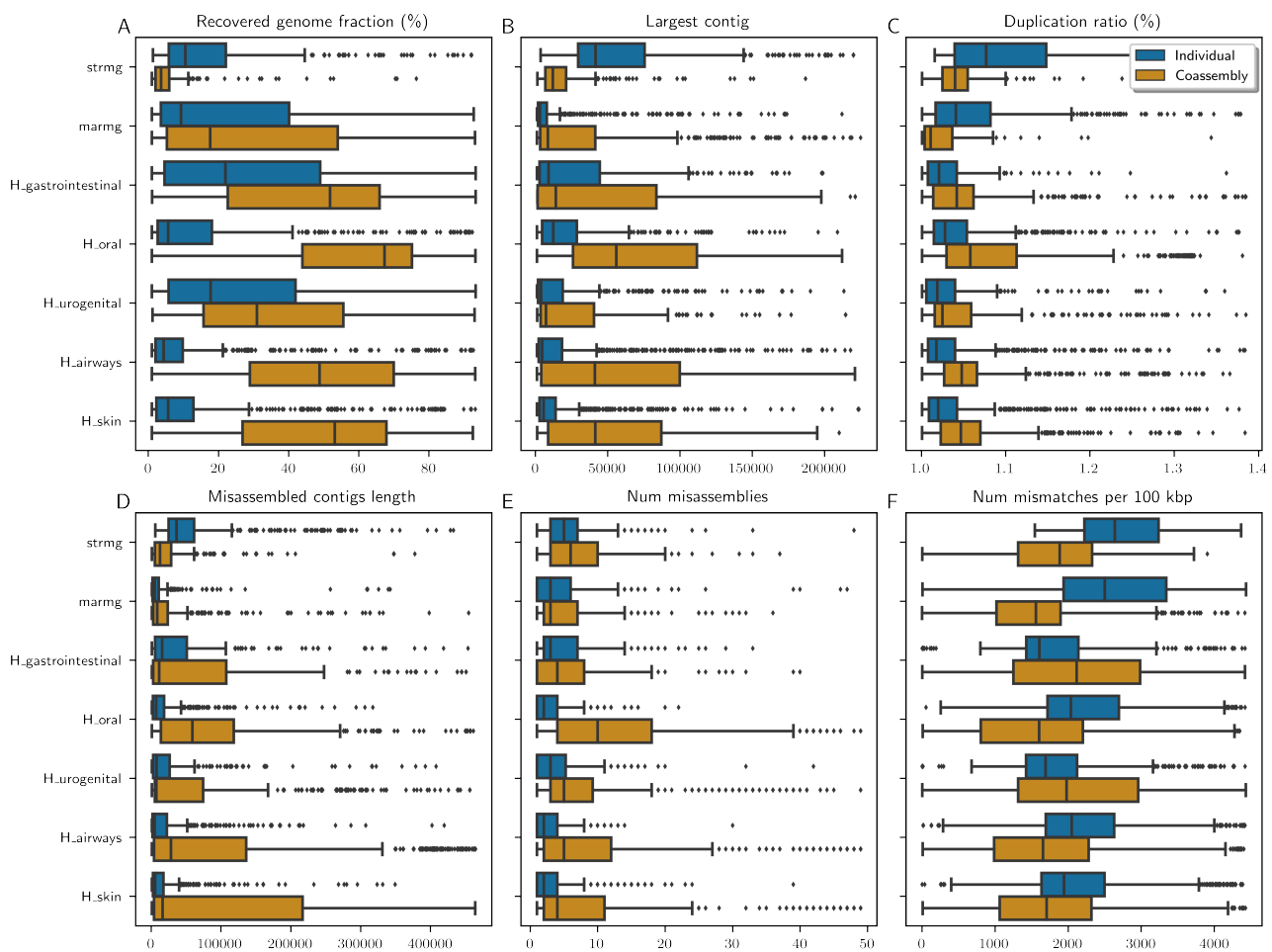
Metaphor would benefit from new third-party software to facilitate the generation of non-prokaryotic bins in the near future. The output of Metaphor's 'annotation' module is suitable for *ad hoc* identification of eukaryotic and viral contigs; after selecting the annotated prokaryotic contigs, it is possible to filter them out, leaving unannotated (putative) eukaryotic and viral contigs. These can then be used as input for a eukaryotic or viral discovery pipeline [50, 51, 52], but this process could be further improved by facilitating the use of custom reference databases in the annotation module. This can also be done directly with the output of the assembly module, but in that case there won't be any screening for prokaryotic contigs. One drawback of this approach is that each eukaryotic/viral discovery pipeline has specific input data formatting requirements. This integration with non-prokaryotic pipelines, along with support for long reads, are priority features to be added to future major versions of Metaphor.

## Availability of Source Code and Requirements

**Project name:** Metaphor

**Project home page:** <https://github.com/vinisalazar/metaphor>

**Documentation:** <https://metaphor-workflow.readthedocs.io/>



**Figure 5. Differences between assembly strategies for each dataset.** Each data point corresponds to a reference genome evaluated with the MetaQUAST tool. Data points above the 98th percentile were classified as outliers and removed from the figure to improve visualisation. See [Figure S1](#) for the full data. The title at the top of each panel indicates the plotted metric. Panels A and C show percentages along the X-axis, while the remainder show absolute values.

**Operating system(s):** Linux, Mac OS (Intel)

**Programming language:** Snakemake (Python 3)

**Other requirements:** Conda, Snakemake v7 or higher, Python 3.7 or higher.

**License:** MIT

**RRID number:** SCR\_023701

### Data availability

This work uses data from the 2nd CAMI Challenge, available from <https://repository.publisso.de/resource/frl:6425521> and <https://repository.publisso.de/resource/frl:6425518>. Analysis code used in this manuscript is available from <https://github.com/vinisalazar/manuscript-notebooks-metaphor>. Snapshots of our code and other data further supporting this work are openly available from the GigaScience repository, GigaDB: Salazar VW; Shaban B; Quiroga M; Turnbull R; Tescari E; Marcelino VR; Verbruggen H; Lê Cao K-A: Supporting data for "Metaphor - A workflow for streamlined assembly and binning of metagenomes" GigaScience Database. 2023. <http://dx.doi.org/10.5524/102408>.

### Declarations

The authors declare they have no competing interests.

### Funding

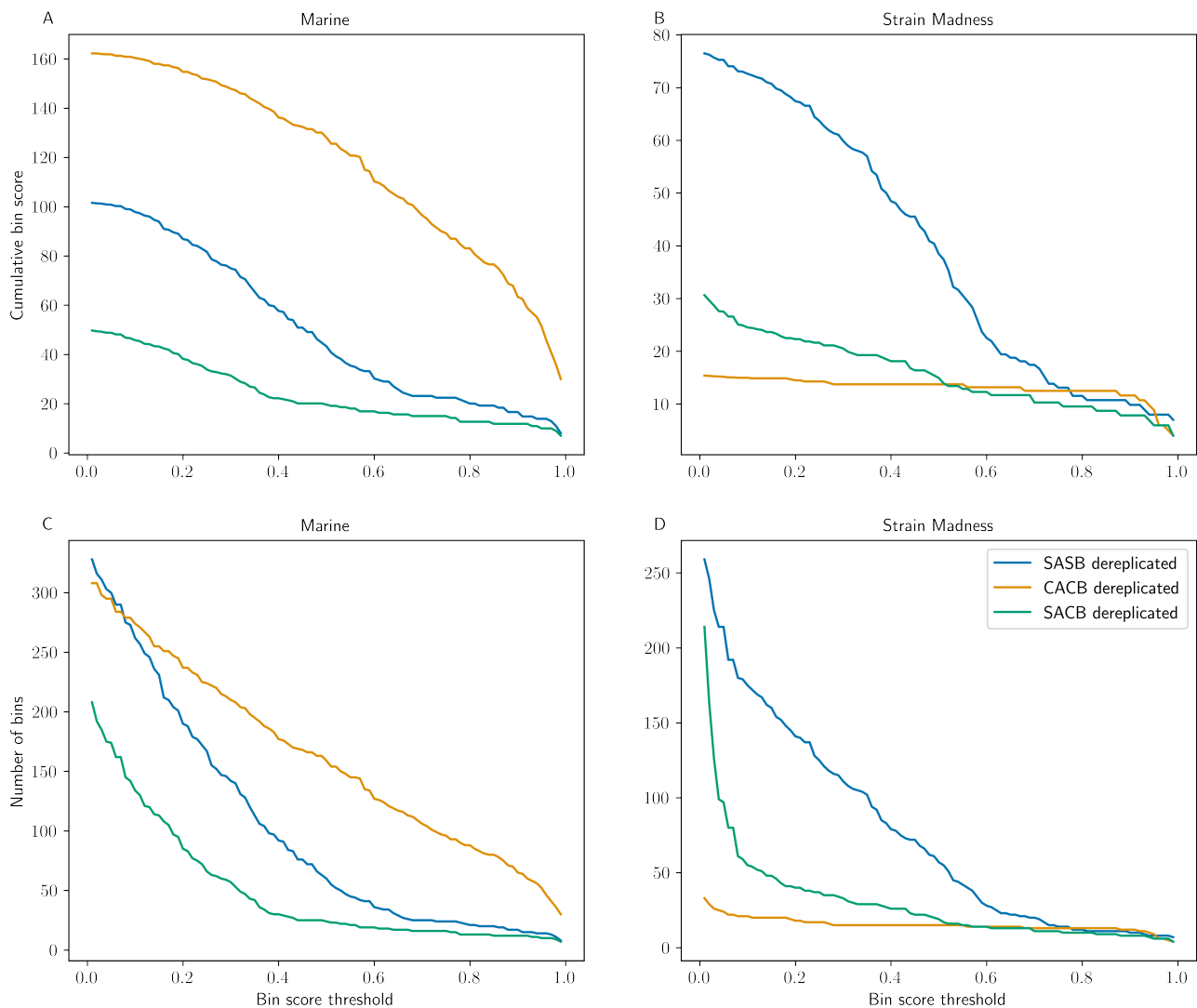
VWS is funded by a Melbourne Research Scholarship from The University of Melbourne. VRM is funded by an Australian Research Council DECRA Fellowship DE220100965. KALC was supported in part by the National Health and Medical Research Council (NHMRC) Career Development fellowship (GNT1159458). This research was also funded by the Australian Research Council project DP200101613.

### Author's Contributions

VWS - Conceptualization, Data curation, Methodology, Investigation Software, Writing - original draft; BS, MMQ, RT, ET - Conceptualization, Writing - review and editing; VRM, HV, KALC - Conceptualization, Supervision, Funding Acquisition, Writing - review and editing.

### Acknowledgments

Metaphor benefited strongly from experience gained developing MetaGenePipe [53], a Cromwell-based workflow for assembly and annotation of metagenomic contigs. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. We thank Francesco Ricci and Uthpala Pushpakumara for providing datasets for early trials



**Figure 6.** Cumulative bin score and number of bins between binning strategies for the Marine and Strain Madness datasets. Lines show the cumulative bin score (A and B) and number of bins (C and D) along the Y-axis, for bins above a certain score threshold (X-axis). Left column shows Marine dataset, and right column shows Strain Madness dataset.

of Metaphor, and colleagues from the Lê Cao lab for sharing their feedback.

## References

- Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* 2021 Jan;39(1):105–114.
- Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 2017;2(11):1533–1542.
- Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* 2018 Jan;5(1):170203.
- Setubal JC. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophysical Reviews* 2021 Dec;13(6):905–909.
- Nelson WC, Tully BJ, Moberley JM. Biases in genome reconstruction from metagenomic data. *PeerJ* 2020 Oct;8:e10119.
- Reiter T, Brooks PT, Irber L, Joslin SEK, Reid CM, Scott C, et al. Streamlining data-intensive biology with workflow systems. *GigaScience* 2021 Jan;10(1).
- Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 2017 Sep;35(9):833–844.
- Meyer F, Fritz A, Deng ZL, Koslicki D, Lesker TR, Gurevich A, et al. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nature Methods* 2022 Apr;19(4):429–440.
- Delgado LF, Andersson AF. Evaluating metagenomic assembly approaches for biome-specific gene catalogues. *Microbiome* 2022 May;10(1):72.
- Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Computational Biology* 2017;
- McMurdie PJ, Holmes S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* 2013 Apr;8(4):e61217.
- Eren AM, Kiehl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi’o. *Nature Microbiology* 2020 Dec;6(1):3–6.
- Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake.



- F1000 Research 2021 Apr;.
14. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018 Sep;34(17):i884–i890.
  15. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. Online resource 2020 Jan;.
  16. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;.
  17. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015 May;31(10):1674–1676.
  18. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016 Apr;32(7):1088–1090.
  19. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018 Sep;34(18):3094–3100.
  20. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *GigaScience* 2021 Feb;10(2):giab008.
  21. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 2010;11:119.
  22. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 2014;.
  23. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* 2021 Jan;49(D1):D274–D281.
  24. Seemann T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014;30(14).
  25. Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* 2021 Jan;.
  26. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;2019(7):1–13.
  27. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nature Methods* 2014 Nov;11(11):1144–1146.
  28. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* 2018 Jul;3(7):836–843.
  29. Churchward B, Millet M, Bihoué A, Fertin G, Chaffron S. MAGNETO: An Automated Workflow for Genome-Resolved Metagenomics. *mSystems* 2022 Jun;0(0):e00432–22.
  30. Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* 2020 Dec;21(1):1–8.
  31. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 2018 Sep;6(1):158.
  32. Krakau S, Straub D, Gourelé H, Gabernet G, Nahnsen S. nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics and Bioinformatics* 2022 Mar;4(1):lqac007.
  33. McKinney W. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 2011;14(9).
  34. Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 2007;9(3):90–95.
  35. Waskom M, Botvinnik O, Ostblom J, Gelbart M, Lukauskas S, Hobson P, et al. Seaborn v0.10.0. Online resource 2020 Apr;.
  36. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation – A benchmark of metagenomics software. *Nature Methods* 2017;14(11):1063–1071.
  37. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods* 2021 Sep;p. 1–8.
  38. Salazar VW. Metaphor’s documentation. Online resource 2023;.
  39. Fritz A, Hofmann P, Majda S, Dahms E, Dröge J, Fiedler J, et al. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 2019 Feb;7(1):17.
  40. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* 2017 Dec;11(12):2864–2868.
  41. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* 2018 Nov;9(1):5114.
  42. Lafayette L, Wiebelt B. Spartan and NEMO: Two HPC-Cloud Hybrid Implementations. 2017 IEEE 13th International Conference on e-Science (e-Science) 2017 Oct;p. 458–459.
  43. 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 Research* 2015;25(7):1043–1055.
  44. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2019;36(6):1925–1927.
  45. Grünig B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* 2018 Jul;15(7):475–476.
  46. Inc A. Conda — Conda documentation. Online resource 2023;.
  47. Jackson M, Kavoussanakis K, Wallace EWJ. Using prototyping to choose a bioinformatics workflow management system. *PLOS Computational Biology* 2021 Feb;17(2):e1008622.
  48. Lai S, Pan S, Sun C, Coelho LP, Chen WH, Zhao XM. metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies. *Genome Biology* 2022 Nov;23(1):242.
  49. Quince C, Nurk S, Raguideau S, James R, Soyer OS, Summers JK, et al. STRONG: metagenomics strain resolution on assembly graphs. *Genome Biology* 2021 Jul;22(1):214.
  50. Pandolfo M, Telatin A, Lazzari G, Adriaenssens EM, Vitulo N. MetaPhage: an Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data. *mSystems* 2022 Sep;7(5):e00741–22.
  51. Karlicki M, Antonowicz S, Karnkowska A. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 2022 Jan;38(2):344–350.
  52. Pronk L, Medema M. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure; 2021.
  53. Shaban B, Quiroga MdM, Turnbull R, Tescari E, Lê Cao KA, Verbruggen H. MetaGenePipe: An Automated, Portable Pipeline for Contig-based Functional and Taxonomic Analysis. *Journal of Open Source Software* 2023 Feb;.

## Supplementary Material

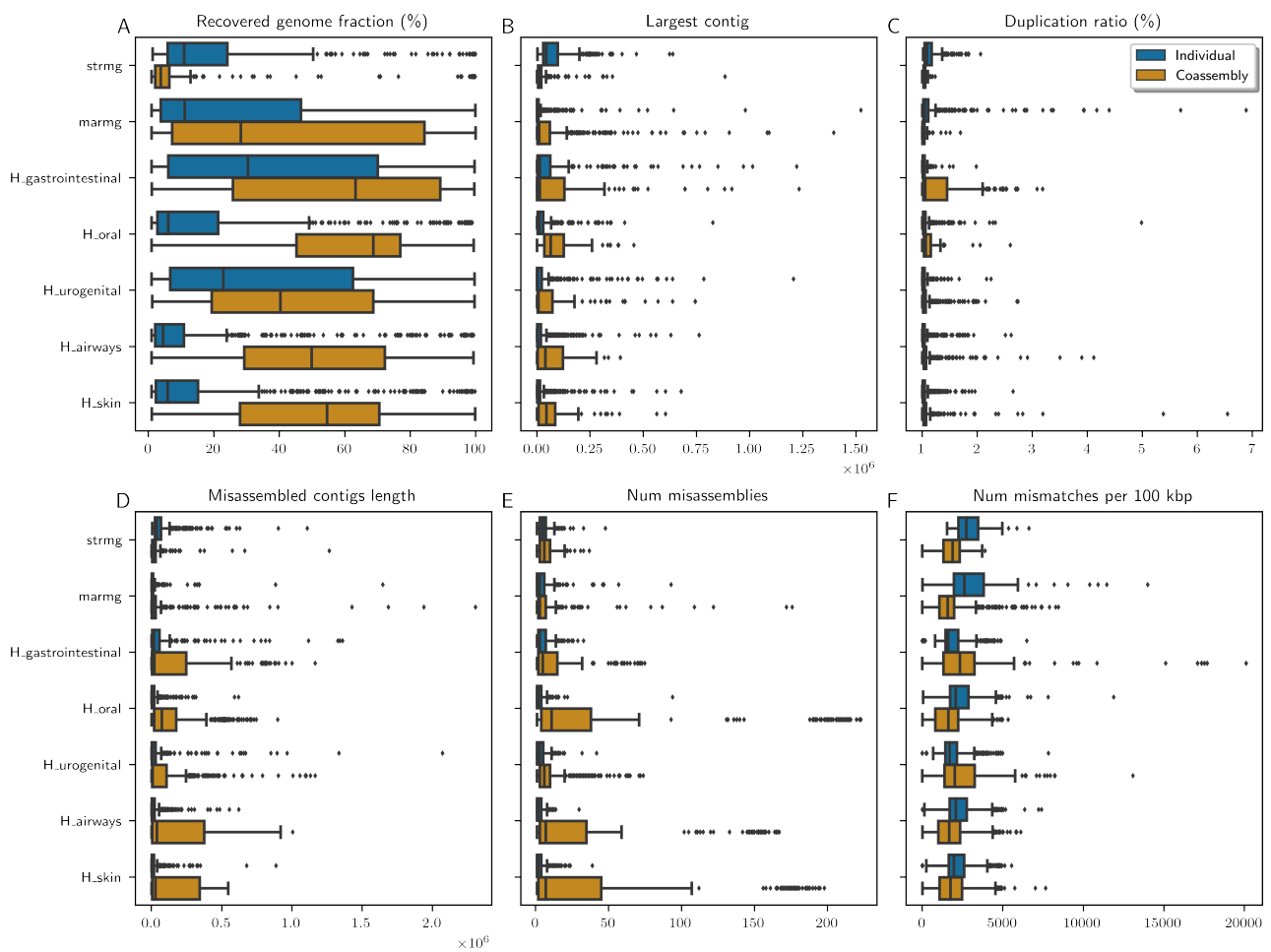


Figure S1. Differences between assembly strategies across datasets. Same data as Fig 5, but including outliers.

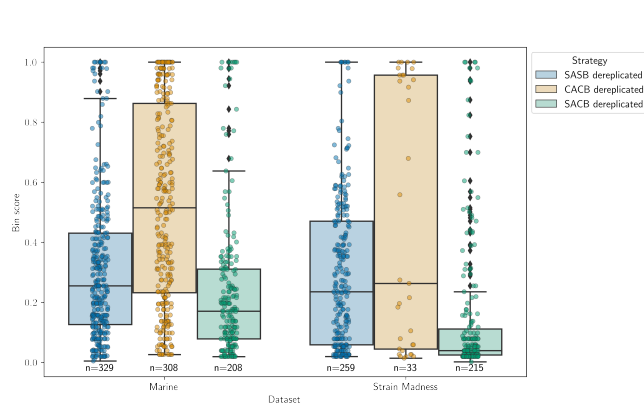


Figure S2. Boxplot of bin scores across different strategies. Each data point is a genome bin, and Y-axis depicts bin scores from 0 to 1. Columns separate datasets, and colours represent different strategies. Numbers underneath each bar show the number of data points for that bar. Bins sets were dereplicated with dRep.

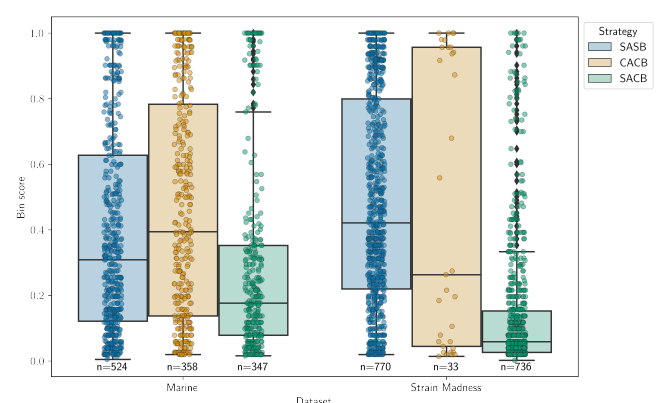
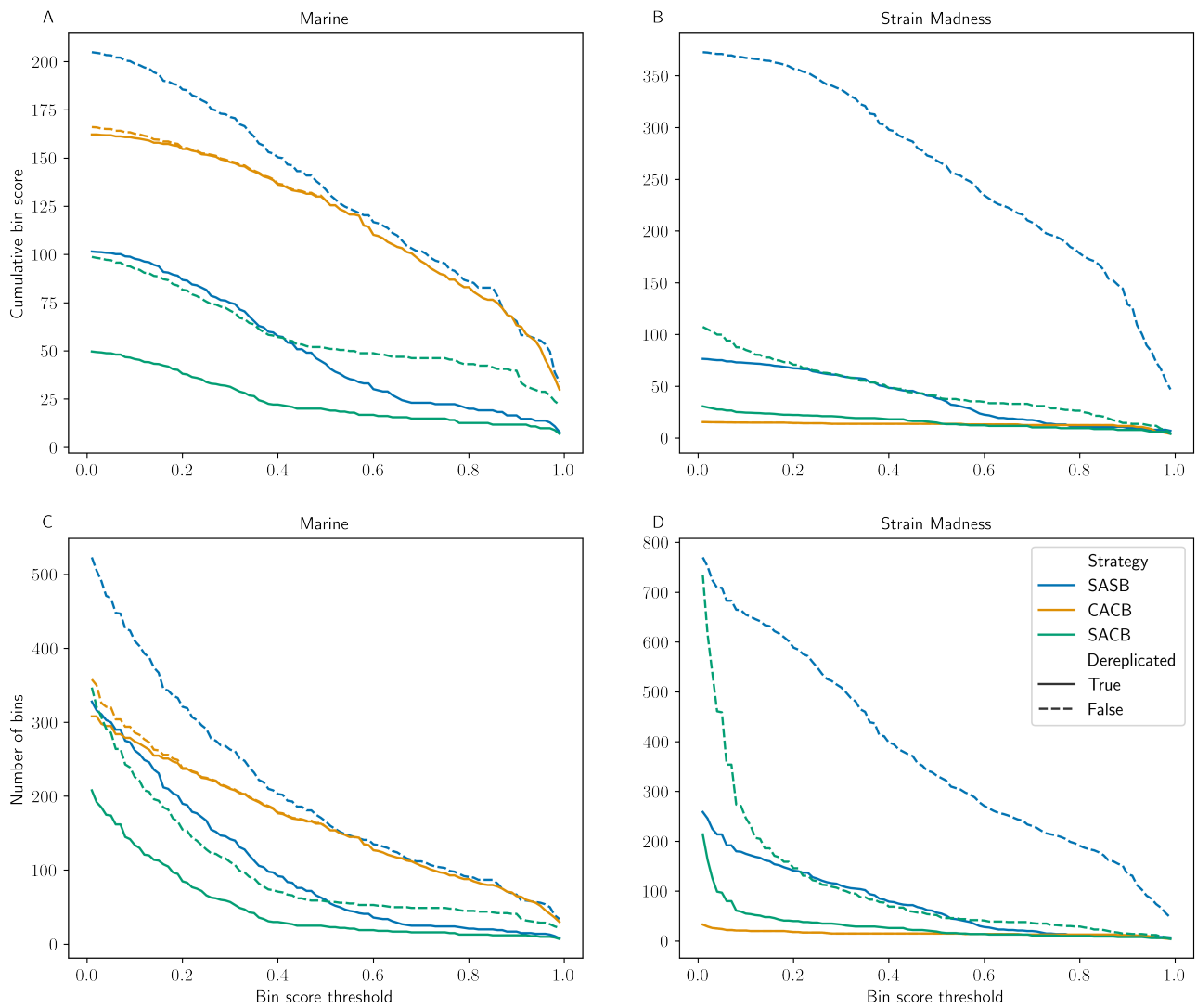


Figure S3. Boxplot of bin scores across different strategies for non-dereplicated data. Each data point is a genome bin, and Y-axis depicts bin scores from 0 to 1. Columns separate datasets, and colours represent different strategies. Numbers underneath each bar show the number of data points for that bar.



**Figure S4.** Cumulative bin score and number of bins between binning strategies for the Marine and Strain Madness datasets. Solid lines show the same data as Fig 6, and dashed lines show data based on bins prior to dereplication with dRep.

**Table S1. Summary of genome bins recovered from the Strain Madness dataset, CACB strategy.** “Bin ID” indicates the binning algorithm that generated the bin, “Bin score  $S_b$ ” is the relative bin score, ‘SCG’ refers to ‘single copy gene’ in “SCG completeness” and “SCG redundancy”, “FastANI reference” and “GTDB classification” refer to the reference genome and corresponding taxonomy assignment. Taxonomy determined with GTDB-Tk v2.3.0, reference data r214 [44].

Bin ID	Bin score $S_b$	SCG completeness	SCG redundancy	FastANI reference	GTDB Classification
metabat2.1221	1	100	0	GCF_004793475.1	Bacteroides sp002491635
concoct.122	1	100	0	GCF_024397795.1	Lactobacillus intestinalis
vamb.S1C5590	1	100	0	GCF_000614185.1	Phocaeicola sartorii
metabat2.4898	1	100	0	GCF_000969835.1	Parabacteroides goldsteinii
concoct.156	0.98039216	98	0	GCF_003030305.1	Cutibacterium acnes
concoct.148	0.97843137	100	2	GCF_001436695.1	Lactobacillus taiwanensis
concoct.92	0.95686275	100	4	GCF_014863545.1	Paenibacillus lautus_A
concoct.121	0.95686275	100	4	GCF_000012845.1	Parabacteroides distasonis
metabat2.328	0.95686275	100	4	GCF_000016825.1	Limosilactobacillus reuteri
vamb.S1C971	0.94117647	94	0	GCF_000392875.1	Enterococcus faecalis
metabat2.3846	0.93678431	98	4	GCA_009911065.1	Ventrimonas sp009911065
concoct.136	0.91668667	96	4	GCF_001027105.1	Staphylococcus aureus
metabat2.1266	0.87258904	96	8	GCF_001544255.1	Enterococcus_B faecium
concoct.115	0.67941176	71	2	GCF_016758115.1	Lactococcus sp002492185
concoct.58	0.55843137	59	2	GCF_013394695.1	Streptococcus sp013394695
metabat2.4512	0.2745098	27	0	GCF_001729805.1	Enterobacter roggenkampii
metabat2.2064	0.26315789	26	0	GCF_000742135.1	Klebsiella pneumoniae
metabat2.1951	0.21568627	22	0	GCF_001457635.1	Streptococcus pneumoniae
metabat2.3969	0.19607843	20	0	GCF_011064845.1	Citrobacter freundii
metabat2.1470	0.18421053	18	0	GCF_000215745.1	Klebsiella aerogenes
concoct.22	0.10526316	11	0	GCF_001729745.1	Enterobacter hormaechei_A
concoct.103_sub	0.07894737	8	0		Unclassified Bacteria
concoct.97_sub	0.05882353	6	0		Citrobacter
concoct.124_sub	0.05882353	6	0		Unclassified Bacteria
concoct.27_sub	0.04473684	16	3		Enterobacter
concoct.91_sub	0.03921569	4	0	GCF_001729745.1	Enterobacter hormaechei_A
concoct.159	0.02631579	3	0		Unclassified Bacteria
concoct.64_sub	0.02631579	3	0		Unclassified Bacteria
vamb.S1C21648	0.02631579	3	0		Unclassified
metabat2.3037_sub	0.01960784	2	0		Unclassified Bacteria
concoct.13	0.01960784	2	0		Unclassified Bacteria
vamb.S1C7072	0.01960784	2	0		Unclassified Bacteria
concoct.35_sub	0.01417112	86	53		Klebsiella



THE UNIVERSITY OF  

---

MELBOURNE

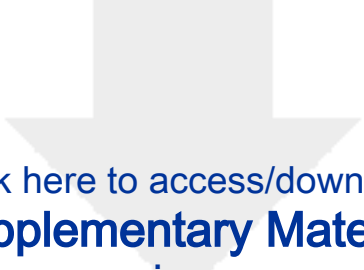
Your PDF file "main.pdf" cannot be opened and processed. Please see the common list of problems, and suggested resolutions below.

Reason:

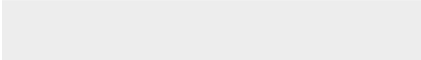

Other Common Problems When Creating a PDF from a PDF file

-----

You will need to convert your PDF file to another format or fix the current PDF file, then re-submit it.



Click here to access/download  
**Supplementary Material**  
main.aux







Click here to access/download  
**Supplementary Material**  
main.blg





Click here to access/download  
**Supplementary Material**  
main.log





Click here to access/download  
**Supplementary Material**  
main.out

