The Journal of Open Source Software

GraphBin-Tk: assembly graph-based metagenomic binning toolkit

Vijini Mallawaarachchi [1¶], **Anuradha Wickramarachchi** [2], **Robert McArthur** [3], **Yapeng Lang** [3], **Katherine Caley** [3], **and Gavin Huttley** [3]

1 Flinders Accelerator for Microbiome Exploration, College of Science and Engineering, Flinders University, Bedford Park, Adelaide, SA 5042, Australia **2** Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Westmead, NSW 2145, Australia **3** Research School of Biology, The Australian National University, Canberra, ACT 2601, Australia ¶ Corresponding author

DOI: 10.21105/joss.07713

Software

- Review
- Repository
- Archive

Editor: Arfon Smith

Reviewers:

- @cumbof
- @vinisalazar
- @telatin

Submitted: 27 September 2024
Published: 28 May 2025

Summary

The study of genetic material directly obtained from natural environments, termed metagenomics, offers valuable insights into microbial communities and their impact on human health and environmental dynamics (Edwards et al., 2013; Pargin et al., 2023). Once the genetic material is extracted, sequenced to obtain reads, and assembled to obtain contigs, a process known as metagenomic binning is used to cluster contigs into bins that represent different taxonomic groups, which results in draft microbial genomes or metagenome-assembled genomes (MAGs) (V. Mallawaarachchi et al., 2024). Several automated metagenomic binning tools incorporating novel computational methods have been introduced (Alneberg et al., 2014; Chandrasiri et al., 2022; D. D. Kang et al., 2019; Pan et al., 2023; Wu et al., 2015; Xue et al., 2022, 2024) which have led to the discovery and characterisation of many novel micro-organisms (Brooks et al., 2017; L. Kang et al., 2024).

Conventional metagenomic binning tools make use of features such as nucleotide composition and abundance information of contigs, yet find it challenging to bin closely related species and contigs with noisy features. Binning tools, such as MetaCoAG (V. Mallawaarachchi & Lin, 2022a, 2022b) that use assembly graphs (which contain the connectivity information) are gaining popularity due to their improved binning results over conventional binning methods. Moreover, assembly graph-based bin refinement tools such as GraphBin (V. Mallawaarachchi et al., 2020) and GraphBin2 (V. G. Mallawaarachchi et al., 2020, 2021) have been introduced to refine binning results from existing binning tools. Yet, these tools exist as individual software packages and running them individually can be complex, time-consuming, and less accessible. Here we present GraphBin-Tk, an assembly graph-based metagenomic binning tool that combines the capabilities of MetaCoAG, GraphBin and GraphBin2, along with additional pre-processing and post-processing functionalities into one comprehensive toolkit. GraphBin-Tk is hosted at https://github.com/metagentools/gbintk.

Statement of need

It is crucial to obtain accurate binning results in metagenomic studies to understand the composition and functional potential of microbial communities. Conventional binning methods mainly rely on two features of contigs; 1) nucleotide composition, represented as normalised frequencies of oligonucleotides (short substrings of a particular length) (Kariin & Burge, 1995) and 2) abundance, the average number of reads that cover each nucleotide base of the contig (Roach et al., 2024; Woodcroft & Newell, 2017). Previous studies have found that these tools face several challenges when binning complex datasets, especially those containing closely related species, short contigs, and shared genomic regions (V. Mallawaarachchi et al., 2024).

To address these challenges, several graph-based metagenomic binning tools (V. Mallawaarachchi et al., 2024) such as MetaCoAG (V. Mallawaarachchi & Lin, 2022a, 2022b) and GraphMB (Lamurias et al., 2022), and bin refinement tools such as GraphBin (V. Mallawaarachchi et al., 2020) and GraphBin2 (V. G. Mallawaarachchi et al., 2020, 2021) have been developed. These tools enhance the binning results by leveraging the connectivity information of the assembly graph – an intermediate output generated from the assembly process, that is often discarded during downstream analysis. However, running these tools individually can be challenging. Users have to install and configure multiple software packages, resolve dependencies and manage different file formats, which increases the risk of errors. Even though existing metagenomic binning toolkits and wrappers such as MetaWRAP (Uritskiy et al., 2018), DAS Tool (Sieber et al., 2018) and MetaBinner (Wang et al., 2023) aim to simplify metagenomic binning workflows, they do not incorporate graph-based binning tools and related processing steps.

GraphBin-Tk addresses the previously mentioned challenges by integrating the capabilities of GraphBin (V. Mallawaarachchi et al., 2020), GraphBin2 (V. G. Mallawaarachchi et al., 2020, 2021) and MetaCoAG (V. Mallawaarachchi & Lin, 2022a, 2022b) in a comprehensive toolkit for metagenomic binning and refinement as shown in Figure 1. Once a metagenome assembly is obtained, GraphBin-Tk enables researchers to bin the assembled contigs and refine bins. Moreover, GraphBin-Tk provides additional functions such as visualisation and evaluation, enabling a wider range of tasks to be performed seamlessly without needing to install and execute additional software. GraphBin-Tk also eliminates any compatibility issues that may arise from running separate binning-related software and enhances the user experience by making the software easier to learn and use. This enables researchers to focus on scientific interpretation, while minimising the technical complexities of using multiple tools.

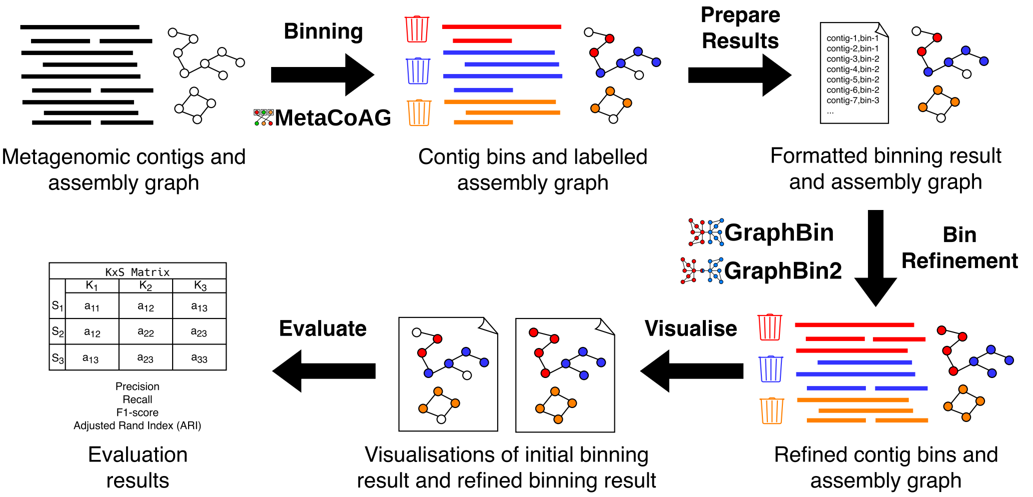

Figure 1: Example binning workflow using tools available from GraphBin-Tk.

Functionality

GraphBin-Tk can perform standalone metagenomic binning using MetaCoAG and bin refinement using either GraphBin or GraphBin2. Additionally, pre- and post-processing functionalities to run these tools and analyse the produced results are included. GraphBin-Tk can be launched using the command `gbintk`. A list of the subcommands provided in GraphBin-Tk is listed in the following table. Further details about the subcommands can be found in the GraphBin-Tk documentation available at https://gbintk.readthedocs.io/.

Subcommand	Tool/processing functionality
`graphbin`	Bin refinement tool GraphBin (V. Mallawaarachchi et al., 2020)
`graphbin2`	Bin refinement tool GraphBin2 (V. G. Mallawaarachchi et al., 2020, 2021)
`metacoag`	Binning tool MetaCoAG (V. Mallawaarachchi & Lin, 2022a, 2022b)
`prepare`	Format initial binning results for GraphBin and GraphBin2
`visualise`	Visualise initial and refined binning results on the assembly graph
`evaluate`	Evaluate binning results given a ground truth

After assembling a metagenomic dataset, a user can start the analysis by running the `metacoag` subcommand to bin the resulting contigs and obtain MAGs as shown in Figure 1. GraphBin-Tk supports metagenome assemblies generated from three popular metagenome assemblers; metaSPAdes (Nurk et al., 2017) and MEGAHIT (Li et al., 2015) for short-read sequencing data and metaFlye (Kolmogorov et al., 2020) for long-read sequencing data. The binning result from MetaCoAG or any other binning tool can be formatted using the `prepare` subcommand into a delimited text file, such as `.csv` or `.tsv`, that represents each contig and its bin name. This formatted binning result can be refined by providing it to either GraphBin or GraphBin2 using the subcommands `graphbin` or `graphbin2`, respectively (Figure 1).

The initial binning result and the refined binning result can be visualised on the assembly graph using the `visualise` subcommand (Figure 1). Users can generate images in different formats such as `png`, `eps`, `pdf`, and `svg`, and customise the dimensions of the images. An example is shown in Figure 2 for the Sim-5G+metaSPAdes dataset (V. G. Mallawaarachchi et al., 2020, 2021) which contains five bacterial species.

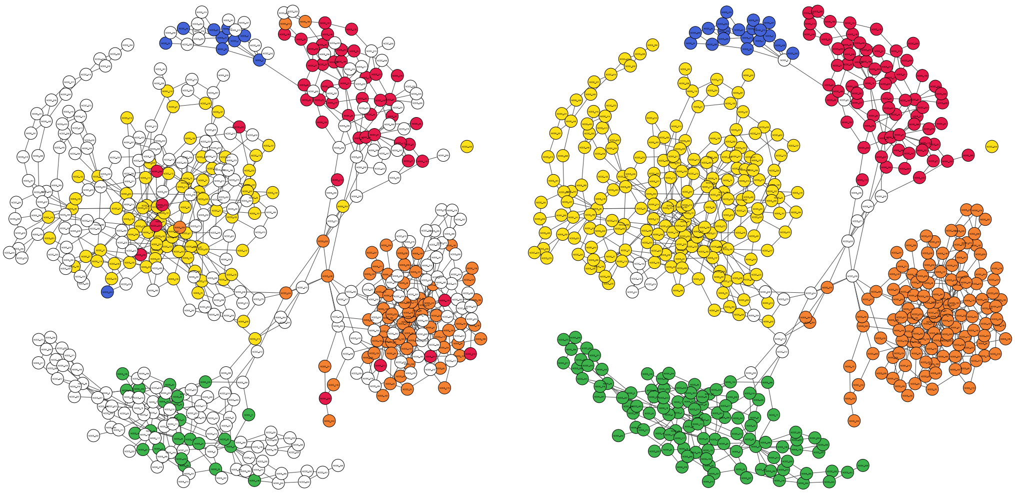

Figure 2: Visualisation of the assembly graph with the initial binning result from MetaCoAG (left) and final binning result from GraphBin (right) for the Sim-5G+metaSPAdes dataset. The vertices represent contigs and the edges represent connections in the assembly graph. The five colours represent the five bins and the white vertices represent unbinned contigs.

Finally, the binning results can be evaluated using the `evaluate` subcommand, by providing the ground truth bins of contigs (Figure 1). This evaluation is only possible for simulated or mock metagenomes where the ground truth genomes of contigs are known. GraphBin-Tk uses the four common metrics 1) precision, 2) recall, 3) F1-score and 4) Adjusted Rand Index (ARI) that have been used in previous binning studies (Alneberg et al., 2014; V. Mallawaarachchi et al., 2020; Meyer et al., 2018). These metrics can be plotted for a comparison between the initial binning result and the refined binning result using custom code. An example is shown in Figure 3 for the Sim-20G+metaSPAdes dataset (V. G. Mallawaarachchi et al., 2020, 2021) containing 20 bacterial species.

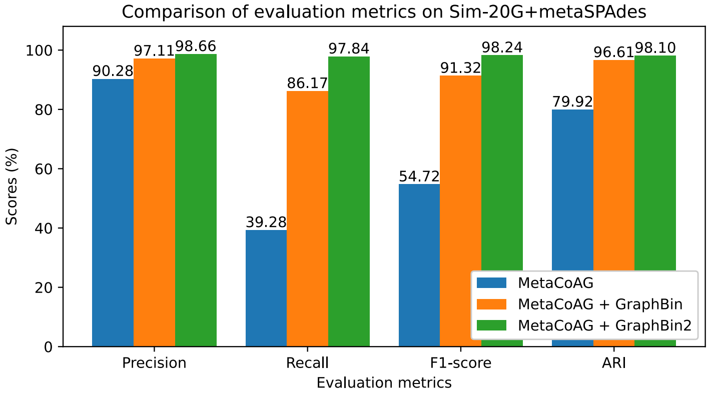

Figure 3: Comparison of evaluation metrics for the initial binning result from MetaCoAG and the refined binning results from GraphBin and GraphBin2 for the Sim-20G+metaSPAdes dataset.

Please refer to the original publications of GraphBin (V. Mallawaarachchi et al., 2020), GraphBin2 (V. G. Mallawaarachchi et al., 2020, 2021) and MetaCoAG (V. Mallawaarachchi & Lin, 2022a, 2022b) for detailed benchmarking results of each tool.

Availability

GraphBin-Tk is distributed as a Conda package, available in the Bioconda channel (Grüning et al., 2018) at https://anaconda.org/bioconda/gbintk. GraphBin-Tk is also available as a Python package on PyPI at https://pypi.org/project/gbintk. The source code is available on GitHub at https://github.com/metagentools/gbintk and features continuous integration, testing coverage, and continuous deployment using GitHub Actions. Detailed documentation and example usage can be found at https://gbintk.readthedocs.io/. The example datasets used for testing purposes are available on Zenodo at https://zenodo.org/records/15313645.

Acknowledgements

This work is dedicated to the memory of the late Dr Yu Lin (The Australian National University) whose guidance and support were instrumental in shaping the original work. His wisdom and mentorship are deeply missed.

This work was supported by an Essential Open Source Software for Science Grant EOSS5-0000000223 from the Chan Zuckerberg Initiative. This work was undertaken with the resources and services from the National Computational Infrastructure (NCI), Australia and Flinders University's Deep Thought high-performance computing platform (Flinders University, 2021).

References

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

Brooks, B., Olm, M. R., Firek, B. A., Baker, R., Thomas, B. C., Morowitz, M. J., & Banfield, J. F. (2017). Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome. *Nature Communications*, *8*(1), 1814. https://doi.org/10.1038/s41467-017-02018-w

Chandrasiri, S., Perera, T., Dilhara, A., Perera, I., & Mallawaarachchi, V. (2022). CH-Bin: A convex hull based approach for binning metagenomic contigs. *Computational Biology and Chemistry*, *100*, 107734. https://doi.org/10.1016/j.compbiolchem.2022.107734

Edwards, R. A., Haggerty, J. M., Cassman, N., Busch, J. C., Aguinaldo, K., Chinta, S., Vaughn, M. H., Morey, R., Harkins, T. T., Teiling, C., Fredrikson, K., & Dinsdale, E. A. (2013). Microbes, metagenomes and marine mammals: Enabling the next generation of scientist to enter the genomic era. *BMC Genomics*, *14*(1), 600. https://doi.org/10.1186/1471-2164-14-600

Flinders University. (2021). *Deep thought (HPC)*. Flinders University. https://doi.org/10.25957/FLINDERS.HPC.DEEPTHOUGHT

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., & Team, T. B. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

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

Kang, L., Song, Y., Mackelprang, R., Zhang, D., Qin, S., Chen, L., Wu, L., Peng, Y., & Yang, Y. (2024). Metagenomic insights into microbial community structure and metabolism in alpine permafrost on the Tibetan Plateau. *Nature Communications*, *15*(1), 5920. https://doi.org/10.1038/s41467-024-50276-2

Kariin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics*, *11*(7), 283–290. https://doi.org/10.1016/S0168-9525(00)89076-9

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T. P. L., & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103–1110. https://doi.org/10.1038/s41592-020-00971-x

Lamurias, A., Sereika, M., Albertsen, M., Hose, K., & Nielsen, T. D. (2022). Metagenomic binning with assembly graph embeddings. *Bioinformatics*, *38*(19), 4481–4487. https://doi.org/10.1093/bioinformatics/btac557

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

Mallawaarachchi, V. G., Wickramarachchi, A. S., & Lin, Y. (2020). GraphBin2: Refined and Overlapped Binning of Metagenomic Contigs Using Assembly Graphs. In C. Kingsford & N. Pisanti (Eds.), *20th international workshop on algorithms in bioinformatics (WABI 2020)* (Vol. 172, pp. 8:1–8:21). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.WABI.2020.8

Mallawaarachchi, V. G., Wickramarachchi, A. S., & Lin, Y. (2021). Improving metagenomic binning results with overlapped bins using assembly graphs. *Algorithms for Molecular Biology*, *16*(1), 3. https://doi.org/10.1186/s13015-021-00185-6

Mallawaarachchi, V., & Lin, Y. (2022a). MetaCoAG: Binning metagenomic contigs via composition, coverage and assembly graphs. In I. Pe'er (Ed.), *Research in computational molecular biology* (pp. 70–85). Springer International Publishing. https://doi.org/10.

1007/978-3-031-04749-7_5

Mallawaarachchi, V., & Lin, Y. (2022b). Accurate Binning of Metagenomic Contigs Using Composition, Coverage, and Assembly Graphs. *Journal of Computational Biology*, *29*(12), 1357–1376. https://doi.org/10.1089/cmb.2022.0262

Mallawaarachchi, V., Wickramarachchi, A., & Lin, Y. (2020). GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics*, *36*(11), 3307–3313. https://doi.org/10.1093/bioinformatics/btaa180

Mallawaarachchi, V., Wickramarachchi, A., Xue, H., Papudeshi, B., Grigson, S. R., Bouras, G., Prahl, R. E., Kaphle, A., Verich, A., Talamantes-Becerra, B., Dinsdale, E. A., & Edwards, R. A. (2024). Solving genomic puzzles: computational methods for metagenomic binning. *Briefings in Bioinformatics*, *25*(5), bbae372. https://doi.org/10.1093/bib/bbae372

Meyer, F., Hofmann, P., Belmann, P., Garrido-Oter, R., Fritz, A., Sczyrba, A., & McHardy, A. C. (2018). AMBER: Assessment of Metagenome BinnERs. *GigaScience*, *7*(6), giy069. https://doi.org/10.1093/gigascience/giy069

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Pan, S., Zhao, X.-M., & Coelho, L. P. (2023). SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics*, *39*(Supplement_1), i21–i29. https://doi.org/10.1093/bioinformatics/btad209

Pargin, E., Roach, M. J., Skye, A., Papudeshi, B., Inglis, L. K., Mallawaarachchi, V., Grigson, S. R., Harker, C., Edwards, R. A., & Giles, S. K. (2023). The human gut virome: Composition, colonization, interactions, and impacts on human health. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.963173

Roach, M. J., Hart, B. J., Beecroft, S. J., Papudeshi, B., Inglis, L. K., Grigson, S. R., Mallawaarachchi, V., Bouras, G., & Edwards, R. A. (2024). Koverage: Read-coverage analysis for massive (meta)genomics datasets. *Journal of Open Source Software*, *9*(94), 6235. https://doi.org/10.21105/joss.06235

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1

Wang, Z., Huang, P., You, R., Sun, F., & Zhu, S. (2023). MetaBinner: A high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biology*, *24*(1), 1. https://doi.org/10.1186/s13059-022-02832-6

Woodcroft, B., & Newell, R. (2017). *WWOOD/coverm: Read coverage calculator for metagenomics*. https://github.com/wwood/CoverM.

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2015). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Xue, H., Mallawaarachchi, V., Xie, L., & Rajan, V. (2024). Encoding Unitig-level Assembly Graphs with Heterophilous Constraints for Metagenomic Contigs Binning. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=vBw8JGBJWj

Xue, H., Mallawaarachchi, V., Zhang, Y., Rajan, V., & Lin, Y. (2022). RepBin: Constraint-Based Graph Representation Learning for Metagenomic Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(4), 4637–4645. https://doi.org/10.1609/aaai.v36i4.20388