




Binette: a fast and accurate bin refinement tool to construct high quality Metagenome Assembled Genomes.

Jean Mainguy ^{1,2,3} and Claire Hoede ^{1,2} ¶

1 Université de Toulouse, INRAE, BioinfOmics, GenoToul Bioinformatics facility, 31326, Castanet-Tolosan, France **2** Université de Toulouse, INRAE, UR 875 MIAT, 31326, Castanet-Tolosan, France **3** LABGeM, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France ¶ Corresponding author

DOI: [10.21105/joss.06782](https://doi.org/10.21105/joss.06782)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Arfon Smith](#) 

Reviewers:

- [@lskatz](#)
- [@beardymcjohnface](#)

Submitted: 31 January 2024

Published: 03 October 2024

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

Statement of need

Metagenomics enables the study of microbial communities and their individual members through shotgun sequencing. An essential phase of metagenomic analysis is the recovery of metagenome-assembled genomes (MAGs). MAGs serve as a gateway to additional analyses, including the exploration of organism-specific metabolic pathways, and form the basis for comprehensive large-scale metagenomic surveys ([Acinas et al., 2021](#); [Nayfach et al., 2019](#)).

In a metagenomic analysis, sequence reads are first assembled into longer sequences called contigs. These contigs are then grouped into bins based on common characteristics in a process called binning to obtain MAGs. There are several tools that can be used to bin contigs into MAGs. These tools are based on various statistical and machine learning methods and use contig characteristics such as tetranucleotide frequencies, GC content and similar abundances across samples ([Alneberg et al., 2014](#); [Kang et al., 2019](#); [Nissen et al., 2021](#)).

The approach of applying multiple binning methods and combining them has proven useful to obtain more and better quality MAGs from metagenomic datasets. This combination process is called bin-refinement and several tools exist to perform such tasks, such as DASTool ([Sieber et al., 2018](#)), MagScot ([Rühlemann et al., 2022](#)) and the bin-refinement module of the metaWRAP pipeline ([Uritskiy et al., 2018](#)). Of these, metaWRAP's bin-refinement tool has demonstrated remarkable efficiency in benchmark analysis ([Meyer et al., 2022](#)). However, it has certain limitations, most notably its inability to integrate more than three binning results. In addition, it repeatedly uses CheckM ([Parks et al., 2015](#)) to assess bin quality throughout its execution, which contributes to its slower performance. Furthermore, since it is embedded in a larger framework, it may present challenges when attempting to integrate it into an independent analysis pipeline.

We present Binette, a bin refinement tool inspired by metaWRAP's bin refinement module, which addresses the limitations of the latter and ensures better results.

Summary

Binette is a Python reimplementation and enhanced version of the bin refinement module used in metaWRAP. It takes as input sets of bins generated by various binning tools. Using these input bin sets, Binette constructs new hybrid bins using basic set operations. Specifically, a bin can be defined as a set of contigs, and when two or more bins share at least one contig, Binette generates new bins based on their intersection, difference, and union ([Figure 1.A](#)).

This approach differs from metaWRAP, which exclusively generates hybrid bins based on bin intersections and allows Binette to expand the range of possible bins.

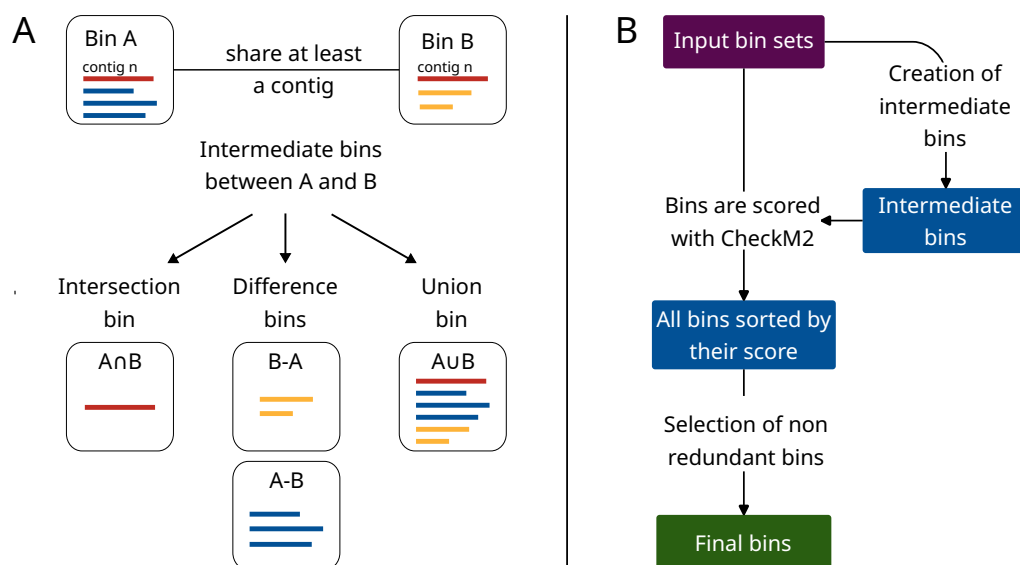


Figure 1: Overview of Binette Steps. (A) Intermediate Bin Creation Example: Bins are represented as square shapes, each containing colored lines representing the contigs they contain. Creation of intermediate bins involves the initial bins sharing at least one contig. Set operations are applied to the contigs within the bins to generate these intermediate bins. **(B) Binette Workflow Overview:** Input bins serve as the basis for generating intermediate bins. Each bin undergoes a scoring process utilizing quality metrics provided by CheckM2. Subsequently, the bins are sorted based on their scores, and a selection process is executed to retain non-redundant bins.

Bin completeness and contamination are assessed using CheckM2 (Chklovski et al., 2023). Bins are scored using the following scoring function: $completeness - weight * contamination$, with the default weight set to 2. These scored bins are then sorted, facilitating the selection of a final new set of non-redundant bins (Figure 1.B). The ability to score bins is based on CheckM2 rather than CheckM1, which is what the metaWRAP pipeline uses. CheckM2 uses a novel approach to evaluate bin quality based on machine learning techniques. This approach improves speed and also provides better results than CheckM1. Binette initiates CheckM2 processing by running its initial steps once for all contigs within the input bins. These initial steps involve gene prediction using Prodigal and alignment against the CheckM2 database using Diamond (Buchfink et al., 2015). Binette uses Pyrodigal (Laralde, 2022), a Python module that uses Cython to provide bindings to Prodigal (Hyatt et al., 2010). The intermediate CheckM2 results are then used to assess the quality of individual bins, eliminating redundant calculations and speeding up the refinement process.

Binette serves as the bin refinement tool within the metagWGS metagenomic analysis pipeline (Mainguy et al., 2024), providing a robust and faster alternative to the bin refinement module of the metaWRAP pipeline as well as other similar bin refinement tools.

Availability

Binette is readily available on PyPI for seamless installation using standard Python package management tools. Additionally, a dedicated Conda package is available in the Bioconda channel (Grüning et al., 2018). The source code for Binette is available on GitHub under the MIT license. The GitHub repository includes continuous integration tests, test coverage, and

employs continuous deployment through GitHub actions to maintain a robust and reliable codebase.

Acknowledgements

We would like to thank Matthias Zytnicki for his valuable insights and support during the development of the binette algorithm.

References

- Acinas, S. G., Sánchez, P., Salazar, G., Cornejo-Castillo, F. M., Sebastián, M., Logares, R., Royo-Llonch, M., Paoli, L., Sunagawa, S., Hingamp, P., Ogata, H., Lima-Mendez, G., Roux, S., González, J. M., Arrieta, J. M., Alam, I. S., Kamau, A., Bowler, C., Raes, J., ... Gasol, J. M. (2021). Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Communications Biology*, 4(1), 1–15. <https://doi.org/10.1038/s42003-021-02112-2>
- Alneberg, J., Bjarnason, B. S., De Bruijn, I., 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>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature Methods*, 20(8), 1203–1212. <https://doi.org/10.1038/s41592-023-01940-w>
- Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., & Team, 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>
- Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11, 1–11. <https://doi.org/10.1186/1471-2105-11-119>
- 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>
- Larralde, M. (2022). Pyrodigal: Python bindings and interface to prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7(72), 4296. <https://doi.org/10.21105/joss.04296>
- Mainguy, J., Vienne, M., Fourquet, J., Darbot, V., Noirot, C., Castinel, A., Combes, S., Gaspin, C., Milan, D., Donnadieu, C., Iampietro, C., Bouchez, O., Pascal, G., & Hoede, C. (2024). metagWGS, a comprehensive workflow to analyze metagenomic data using illumina or PacBio HiFi reads. *bioRxiv*. <https://doi.org/10.1101/2024.09.13.612854>
- Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., & others. (2022). Critical assessment of metagenome interpretation: The second round of challenges. *Nature Methods*, 19(4), 429–440. <https://doi.org/10.1038/s41592-022-01431-4>
- Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., & Kyrpides, N. C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753),

- 505–510. <https://doi.org/10.1038/s41586-019-1058-x>
- Nissen, J. N., Johansen, J., Allesøe, R. L., Søndersby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & others. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, 39(5), 555–560. <https://doi.org/10.1038/s41587-020-00777-4>
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), 1043–1055. <https://doi.org/10.1101/gr.186072.114>
- Rühlemann, M. C., Wacker, E. M., Ellinghaus, D., & Franke, A. (2022). MAGScoT: A fast, lightweight and accurate bin-refinement tool. *Bioinformatics*, 38(24), 5430–5433. <https://doi.org/10.1093/bioinformatics/btac694>
- Sieber, C. M., 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), 1–13. <https://doi.org/10.1186/s40168-018-0541-1>