A Review of the Advancements in Long-Read Metagenomic Binning Over the Years

1st Nethmi Ranasinghe
Department of Computer Engineering
University of Peradeniya
Sri Lanka
e18282@eng.pdn.ac.lk

2nd Sathsarani Aththanayaka
Department of Computer Engineering
University of Peradeniya
Sri Lanka
e18030@eng.pdn.ac.lk

3rd Jayathri Ranasinghe
Department of Computer Engineering
University of Peradeniya
Sri Lanka
e18283@eng.pdn.ac.lk

4th Dr. Damayanthi Herath
Department of Computer Engineering
University of Peradeniya
Sri Lanka
damayanthiherath@eng.pdn.ac.lk

5th Dr. Vijini Mallawarachchi
Bioinformatics
Flinders University
Australia
vijini.mallawaarachchi@flinders.edu.au

***Abstract*—The study of microbial communities has undergone significant advancements with the advent of long-read sequencing which offered significant improvements when compared to short-read sequencing. In recent years, a multitude of long-read binning tools has emerged, categorizing themselves as either reference-based or reference-free. Notably, reference-free binning has gained popularity due to its capacity to discover novel species. These long reads binning tools utilize different unsupervised learning approaches, using coverage and composition features of long reads. This review intends to showcase various long-read binning tools by emphasizing their distinctive features and the methodologies employed in each tool, including the binning or clustering algorithms utilized. Additionally, we will conduct a comprehensive comparison of the tools by examining their performance with specific datasets. The discussion extends to the role of refiners in enhancing binning accuracy. Overall, the paper outlines current findings and proposes directions for future research in this dynamic field.**

***Index Terms*—Metagenomics binning, Long reads, Machine learning, Clustering, Bin Refiners**

I. Introduction

Microorganisms live in diverse environments across the Earth, playing crucial roles in human health, agriculture, food production, climate changes, and various other processes [1]. Every living organism is constructed from tiny units called cells which serve two crucial purposes: structure and function. Within the nucleus of each cell lies the genome, a complete blueprint containing the instructions for building and maintaining the entire organism, including its unique characteristics and behaviors. This blueprint is housed within thin, thread-like structures known as chromosomes, which are composed of DNA and proteins. DNA, the molecule carrying the genetic instructions, takes the form of a double helix, two long strands twisted together. It is made up of repeating units called nucleotides, each labeled with a specific letter: A(Adenine), C(Cytosine), G(Guanine), or T(Thymine). Genes, the fundamental units of heredity, are segments of DNA that contain the instructions for building proteins or functional RNA molecules. They act as messengers, passing on traits from one generation to the next, ensuring the continuity of life [2].

Metagenomics involves studying the genetic material of microorganisms directly from their natural environment, such as soil, the gut, the ocean, and more. This approach eliminates the necessity for laboratory culturing which can introduce biases in the culturing process. Also, it leads to the discovery of vast new lineages of microbial life [3].

First and next-generation sequencing(e.g. Illumina) technologies produce short reads whereas it is necessary to assemble these short reads into contigs that have richer information for binning (e.g.CoMet [4], MetaCOAG [5], MetaBAT [6], BM3C3 [7], VAMB [8]). However, short reads encounter challenges when dealing with repeated or similar sequences in the DNA, making it harder to assemble a complete and accurate representation of the genome [9].

Third-generation sequencing technologies such as Pacific Bioscience (PacBio) [10] and Oxford Nanopore (ONT) entered the spotlight by introducing long reads which are much longer than short reads(less than 10kbp). This increased length eliminates the need for contigs.

Metagenomic binning is an important area of metagenomic studies that facilitates the grouping of sequences into taxonomic groups to reconstruct microbial genomes. Mainly there are two methods in binning; 1) Reference-based (supervised) binning and 2) Reference-free (unsupervised) binning. Reference-based binning(e.g. Megan-LR [11], Kraken [12], Kaiju [9]) adopts a way of binning that compares similarities of sequences with respect to a reference database of the known genome. However, a drawback of this method is the limited availability of reference databases [13].

Conversely, reference-free binning(e.g. MetaBCC-LR [14], LRBinner [15], OBLR [16]) does not rely on a reference database. Instead, it uses computational methods and tries to

group reads based on read qualities so that reads with the same species are clustered together. This approach is particularly well-suited for the identification of novel or rare species.

However, directly applying contig binning tools to classify long reads proved unfeasible, primarily due to the absence of coverage information for individual long reads. Additionally, raw long-read datasets are more extensive in size compared to the typical datasets containing assembled contigs [15]. Recognizing these limitations, researchers have turned their attention to developing tailored strategies to address the unique characteristics of long reads.

II. Long-reads binning tools

A. Overview

In the literature around 2017, a significant surge was noted in the creation of specialized tools designed for the binning of long reads. Among these, Megan-LR [11] stands out as one of the earliest tools, employing a reference database. Megan-LR utilizes a protein-alignment-based approach and introduces two algorithms; one for taxonomic binning (based on Lowest Common Ancestor) and another for functional binning (based on an Interval-tree algorithm).

Two other noteworthy reference-independent tools, MetaProb [17] and BusyBee Web [18], significantly contributed to the domain of unsupervised metagenomic binning. BusyBee Web, in particular, includes a web-based interface, offering additional visual insights into the binning process. On the other hand, MetaProb introduced a novel approach called probabilistic sequence signature, which proved to be a notable advancement in the field. However, despite their respective strengths, both MetaProb and BusyBee Web faced challenges related to scalability as input dataset sizes increased, impeding their ability to bin entire datasets in a single iteration.

To address these scalability issues, MetaBCC-LR [14] was introduced, featuring a novel approach to represent the abundance of long reads. It surpassed the limitations of its predecessors, which solely relied on the composition feature, thereby achieving higher accuracy. Subsequently, more advanced tools, such as LRBinner [15], OBLR [16], and SemiBin2 [19], emerged. These tools employed supervised learning techniques such as neural networks, to improve the accuracy and efficiency of the overall process which will be in-depth discussed as we progress through the review.

It is important to mention that the landscape of long-read metagenomic binning tools remains relatively limited, reflecting the novelty of this technology in the current world.

B. Read features and feature extraction strategies

In the context of the aforementioned reference-free binning tools, the majority employ composition and coverage as the read features for the binning process. Composition is the relative abundance of distinct short sequences called oligonucleotides within the reads. This is observed to be conserved within a given species and distinct between species [20], [21] Simultaneously, coverage, representing the count of reads covering a specific region of an underlying genome, is crucial in metagenomic binning. Long reads from the same species typically exhibit similar coverages [8], [14]. These features should be represented as numerical feature vectors to facilitate computational analyses.

The commonly used method for determining coverage features involves k-mer coverage histograms, often with a relatively large value for k. Notably, MetaBCC-LR and LRBinner employ 15-mers to generate coverage histograms for individual reads. While the k-mer-based approach yields promising results in long-read binning, it is susceptible to unreliable coverage estimation for individual long reads and exhibits poor sensitivity for low-abundance species due to imbalanced clusters [15].

In response to these limitations, the latest tool, OBLR, adopts an alternative approach with read overlap graphs to estimate read coverages, resulting in improved binning outcomes with elevated accuracy. In this approach, the node degree is used to estimate the coverage of the corresponding read [16].

The computation of the composition feature is executed through the analysis of oligonucleotide frequency profiles in all three tools. Specifically, MetaBCC-LR and LRBinner utilize trinucleotide frequency vectors, while OBLR employs tetranucleotide frequency vectors for each read.

Although MetaBCC-LR and LRBinner share a common methodology for computing their feature vectors, the main difference lies in their approach to the binning process. MetaBCC-LR utilizes the coverage information of reads to initially cluster them, followed by a secondary binning process using composition information. Notably, only a subset of reads from the entire dataset is employed for this procedure. Toward the end, statistical models are crafted for each identified bin to cluster the remaining reads [14].

In contrast, LRBinner simultaneously computes composition and coverage information for the entire dataset and merges them through a variational autoencoder. This innovative approach addresses challenges faced by MetaBCC-LR, especially in accurately binning species with non-uniform composition or coverage and avoiding the misclassification of species with low abundance as non-genomic. Furthermore, the overall binning accuracy is enhanced as LRBinner eliminates the need to subsample large datasets [15].

Despite high accuracy levels, LRBinner still suffers from the challenge of distinguishing long reads from similar regions shared between different species as it does not support overlapped binning. Also when it comes to the process of assembly, the possibility of introducing more fragmented assemblies is stated as a potential limitation [15].

In the case of SemiBin2, tetramer frequencies and preprocessed abundance values of each read are employed, passing through a self-supervised deep learning model [19]. Here, abundance serves as a feature similar to coverage, measured using a tool called BEDTools [22].

C. Clustering algorithms

In reference-free binning tools, the methods employed for clustering typically involve unsupervised machine learning-based approaches.

It has been observed that these tools prefer density-based clustering algorithms over traditional centroid-based clustering methods such as k-means. Unlike traditional centroid-based algorithms like k-means, which assume clusters to have well-defined centers, these tools prefer density-based clustering due to the arbitrary shapes and sparse regions of read clusters.

For instance, MetaBCC-LR employs DBSCAN, a density-based clustering algorithm, to effectively group sampled reads. This algorithm needs a user-tunable parameter (ϵ) to define the maximum distance at which two points are considered to be connected [23]. SemiBin2 uses an ensemble-based DBSCAN approach with different ϵ values in each model. LRBinner takes a different approach, utilizing its own distance-based grouping algorithm, which uses the latent space generated by the Variational Autoencoder. Lastly, OBLR utilizes HDBSCAN, a hierarchical density-based clustering algorithm [24], to cluster reads in a more advanced manner.

Metagenomics samples are observed to have imbalanced clusters as they consist of species with varying coverages. This can lead to the well-known class imbalance problem when the entire dataset is clustered at once [25]. Therefore, an important step of clustering in the OBLR tool is its sub-sampling strategy to address the above issue. It uses a probabilistic down sampling approach which has resulted in clusters with similar sizes and less isolated points when compared with uniform sampling. After clustering the selected sample of reads using the HDBSCAN, it uses inductive learning to effectively assign bins to the remaining reads. Rather than using classical label propagation techniques which are less scalable and inefficient for large-scale graphs, it employs the GraphSAGE neural network architecture [16].

TABLE I
A COMPARISON OF THE LONG-READ METAGENOMIC BINNING TOOLS

Tools	**Read features considered**		**Clustering algorithms**
	Composition	**Coverage**	
MetaBCC-LR [14]	Trinucleotide frequency vector	15-mer coverage histogram	DBSCAN
LRBinner [15]	Trinucleotide composition vector	15-mer coverage vector	Distance-based clustering
OBLR [16]	Tetranucleotide frequency vector	Node degree of the read overlap graph	HDBSCAN
SemiBin2 [19]	Tetramer frequency	Estimated abundance	Ensemble based DBSCAN

D. Binning performance comparison

In the comparison of accuracy for binning results against MetaBCC-LR, LRBinner, and OBLR, the focus is primarily on precision, recall, and F1-score and the number of bins produced across different datasets. The evaluation utilizes datasets categorized into simulated data and real data. SimLoRD [26], a read simulator for third-generation sequencing with a focus on the Pacific Biosciences SMRT error model, is employed to simulate four PacBio datasets. These datasets, named Sim-8, Sim-20, Sim-50, and Sim-100, consist of 8, 20, 50, and 100 species, respectively, with an average read length of 5000 bp [15], [16]. The real datasets used in the evaluation include ZymoEVEN(Oxford Nanopore reads), SRR9202034 (PacBio CCS reads), and SRX9569057 (PacBio-HiFi reads). These real datasets are chosen for their known ground truth references [15], [16]. According to the comparative evaluation done in the OBLR paper, it is evident that OBLR exhibits better accuracy in binning compared to the other two tools.

During the execution of these tools, input files can be provided in the FASTA or FASTQ format, and the output directory will contain the final output files. These output files include a text file containing read IDs and their corresponding bins, along with log files and checkpoints. After the completion of the binning procedure, the binned reads are assembled using long-read assemblers such as wtdbg2 [27] and metaFlye [**?**]. Notably, the prior binning procedure has significantly reduced peak-memory usage. While LRBinner shows limited improvement in CPU time after reads are binned, OBLR exhibits more substantial improvement. It is challenging to perform a direct comparison of CPU time values due to the varied computing setups used in the experiments of these tools.

III. IMPROVED BINNING THROUGH BIN REFINERS

The performance of metagenomic binning tools has raised concerns, particularly in handling complex microbial communities, as their outcomes may not always be satisfactory. Due to different algorithms or statistical models used in such tools, inconsistencies can be seen in the binning results across different binning tools. After the binning process, ensuring that all sequences within an identified bin are exclusive to a particular species and free of contamination from other species is crucial to prevent misleading conclusions [29].

In that case, to enhance the precision of binned reads, the use of bin refiners becomes essential. Notable examples include Binning refiner [29], d2sbin [30], GraphBin [31], UGMAGrefiner [32], and METAMVGL [33]. Many of these refining tools rely on the additional information derived from assembly graphs which is a prime element in contig binning. Consequently, they are suitable for refining results obtained from contig binning tools. However, the existing literature does not provide sufficient information regarding bin refiners specifically designed for long-read metagenomic binning tools.

IV. CONCLUSION AND FUTURE PERSPECTIVE

In conclusion, the emergence of long-read metagenomics represents a significant step forward in genomic analysis, providing a more comprehensive view of microbial communities. Despite being a recent advancement, a significant number of long-read binning tools have been developed, with an

admirable level of accuracy. However, our literature search revealed a lack of diversity among these tools compared to their short-read counterparts. A notable reliance on composition and coverage as read features is shown across the majority of existing long-read binning tools.

This observation calls for further exploration and innovation in the identification of additional read features that could enhance the accuracy and refine the results of long-read binning tools. Additionally, we have observed that these tools are not primarily designed with a focus on multi-kingdoms of microorganisms. Exploring this domain more comprehensively and seeking a more sophisticated approach to incorporate kingdom-level information into the binning process could potentially enhance the results.

Given the considerable size of long-read datasets, it is also important to address efficiency concerns. Optimizing computational efficiency in long-read binning tools is crucial for handling such massive volumes of data generated by these sequencing technologies.

Moreover, the advanced error correction mechanisms have resulted in an accuracy of 99% in the long reads sequenced data [34] showing the potential to do further enhancements in the field.

REFERENCES

[1] Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., Behrenfeld, M. J., Boetius, A., Boyd, P. W., Classen, A. T., Crowther, T. W., Danovaro, R., Foreman, C. M., Huisman, J., Hutchins, D. A., Jansson, J. K., Karl, D. M., Koskella, B., Mark Welch, D. B., Martiny, J. B. H., ... Webster, N. S. (2019). Scientists' warning to humanity: microorganisms and climate change. Nature reviews. Microbiology, 17(9), 569–586. https://doi.org/10.1038/s41579-019-0222-5

[2] Ray, S. (2014). The Cell: A Molecular Approach. The Yale Journal of Biology and Medicine, 87(4), 603–604.

[3] Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiology and molecular biology reviews : MMBR, 68(4), 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

[4] Herath, D., Tang, S. L., Tandon, K., Ackland, D., & Halgamuge, S. K. (2017). CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. BMC bioinformatics, 18(Suppl 16), 571. https://doi.org/10.1186/s12859-017-1967-3

[5] Mallawaarachchi, V., & Lin, Y. (2022). Accurate Binning of Metagenomic Contigs Using Composition, Coverage, and Assembly Graphs. Journal of computational biology : a journal of computational molecular cell biology, 29(12), 1357–1376. https://doi.org/10.1089/cmb.2022.0262

[6] Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ, 3, e1165. https://doi.org/10.7717/peerj.1165

[7] Yu, G., Jiang, Y., Wang, J., Zhang, H., & Luo, H. (2018). BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. Bioinformatics, 34, 4172–4179. https://doi.org/10.1093/bioinformatics/bty519

[8] Nissen, J.N., Johansen, J., Allesøe, R.L. et al. (2021). Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 39, 555–560 . https://doi.org/10.1038/s41587-020-00777-4

[9] Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature communications, 7, 11257. https://doi.org/10.1038/ncomms11257

[10] Xie, H., Yang, C., Sun, Y., Igarashi, Y., Jin, T., & Luo, F. (2020). PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning. Frontiers in genetics, 11, 516269. https://doi.org/10.3389/fgene.2020.516269

[11] Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D., & Williams, R. B. H. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biology direct, 13(1), 6. https://doi.org/10.1186/s13062-018-0208-7

[12] Wood, D.E., Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15, R46 . https://doi.org/10.1186/gb-2014-15-3-r46

[13] Madival, S. D., Mishra, D. C., Sharma, A., Kumar, S., Maji, A. K., Budhlakoti, N., Sinha, D., & Rai, A. (2022). A Deep Clustering-based Novel Approach for Binning of Metagenomics Data. Current genomics, 23(5), 353–368. https://doi.org/10.2174/1389202923666220928150100

[14] Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., & Lin, Y. (2020). MetaBCC-LR: metagenomics binning by coverage and composition for long reads. Bioinformatics (Oxford, England), 36(Suppl_1), i3–i11. https://doi.org/10.1093/bioinformatics/btaa441

[15] Wickramarachchi, A., & Lin, Y. (2022). Binning long reads in metagenomics datasets using composition and coverage information. Algorithms for molecular biology : AMB, 17(1), 14. https://doi.org/10.1186/s13015-022-00221-z

[16] Wickramarachchi, A., & Lin, Y. (2022, May). Metagenomics binning of long reads using read-overlap graphs. In RECOMB International Workshop on Comparative Genomics (pp. 260-278). Cham: Springer International Publishing.

[17] Girotto, S., Pizzi, C., & Comin, M. (2016). MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. Bioinformatics (Oxford, England), 32(17), i567–i575. https://doi.org/10.1093/bioinformatics/btw466

[18] Laczny, C. C., Kiefer, C., Galata, V., Fehlmann, T., Backes, C., & Keller, A. (2017). BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. Nucleic acids research, 45(W1), W171–W179. https://doi.org/10.1093/nar/gkx348

[19] 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 (Oxford, England), 39(39 Suppl 1), i21–i29. https://doi.org/10.1093/bioinformatics/btad209

[20] Strous, M., Kraft, B., Bisdorf, R., & Tegetmeyer, H. E. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. Frontiers in microbiology, 3, 410. https://doi.org/10.3389/fmicb.2012.00410

[21] Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics (Oxford, England), 32(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

[22] Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

[23] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

[24] McInnes, Leland & Healy, J30ohn & Astels, Steve. (2017). hdbscan: Hierarchical density based clustering. The Journal of Open Source Software. 2. 10.21105/joss.00205.

[25] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

[26] Stöcker, B. K., Köster, J., & Rahmann, S. (2016). SimLoRD: Simulation of Long Read Data. Bioinformatics (Oxford, England), 32(17), 2704–2706. https://doi.org/10.1093/bioinformatics/btw286

[27] Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nature methods, 17(2), 155–158. https://doi.org/10.1038/s41592-019-0669-3.

[28] 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

[29] Song, W. Z., & Thomas, T. (2017). Binning_refiner: improving genome bins through the combination of different binning programs. Bioinformatics (Oxford, England), 33(12), 1873–1875. https://doi.org/10.1093/bioinformatics/btx086

[30] Wang, Y., Wang, K., Lu, Y. Y., & Sun, F. (2017). Improving contig binning of metagenomic data using d2S oligonucleotide frequency dissimilarity. BMC bioinformatics, 18(1), 425. https://doi.org/10.1186/s12859-017-1835-1

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

[32] Xiang, B., Zhao, L., & Zhang, M. (2023). Unitig level assembly graph based metagenome-assembled genome refiner (UGMAGrefiner): A tool to increase completeness and resolution of metagenome-assembled genomes. Computational and structural biotechnology journal, 21, 2394–2404. https://doi.org/10.1016/j.csbj.2023.03.030

[33] Zhang, Z., & Zhang, L. (2021). METAMVGL: a multi-view graph-based metagenomic contig binning algorithm by integrating assembly and paired-end graphs. BMC bioinformatics, 22(Suppl 10), 378. https://doi.org/10.1186/s12859-021-04284-4

[34] Zhang, H., Jain, C., & Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. BMC genomics, 21(Suppl 6), 889. https://doi.org/10.1186/s12864-020-07227-0