Genome analysis

Supplementary Material for "MetaBCC-LR: Metagenomics Binning by Coverage and Composition for Long Reads"

Anuradha Wickramarachchi [1,*], Vijini Mallawaarachchi [1], Vaibhav Rajan [2] and Yu Lin [1,*]

[1] Research School of Computer Science, Australian National University, Canberra, ACT 0200, Australia and
[2] Department of Information Systems and Analytics, School of Computing, National University of Singapore, Singapore 117417.

*To whom correspondence should be addressed.

Abstract

Motivation: Metagenomics studies have provided key insights into the composition and structure of microbial communities found in different environments. Among the techniques used to analyse metagenomic data, binning is considered a crucial step to characterise the different species of microorganisms present. The use of short-read data in most binning tools poses several limitations, such as insufficient species-specific signal, and the emergence of long-read sequencing technologies offers us opportunities to surmount them. However, most current metagenomic binning tools have been developed for short reads. The few tools that can process long reads either do not scale with increasing input size or require a database with reference genomes that are often unknown. In this paper, we present MetaBCC-LR, a scalable reference-free binning method which clusters long reads directly based on their k-mer coverage histograms and oligonucleotide composition.

Results: We evaluate MetaBCC-LR on multiple simulated and real metagenomic long-read datasets with varying coverages and error rates. Our experiments demonstrate that MetaBCC-LR substantially outperforms state-of-the-art reference-free binning tools, achieving ∼13% improvement in F1-score and ∼30% improvement in ARI compared to the best previous tools. Moreover, we show that using MetaBCC-LR before long read assembly helps to enhance the assembly quality while significantly reducing the assembly cost in terms of time and memory usage. The efficiency and accuracy of MetaBCC-LR pave the way for more effective long-read based metagenomics analyses to support a wide range of applications.

Availability: The source code is freely available at: https://github.com/anuradhawick/MetaBCC-LR.

Contact: anuradha.wickramarachchi@anu.edu.au and yu.lin@anu.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Trinucleotide Composition and k-mer Coverage Distributions of ONT Reads

Figure 1 denotes the Trinucleotide composition of 100 non-overlapping Oxford Nanopore (ONT) reads simulated from the reference genome of *P. aeruginosa*. We can see that the trinucleotide frequencies of ONT reads follow a close pattern to that of the reference genome despite the high error rates.

2 Datasets

Detailed information about the simulated PacBio datasets such as the type of reads simulated, the species present, size of the genomes of each species, coverage, size of the dataset and average read length can be found in Table 1. Details about the simulated Nanopore (ONT) datasets can be found in Table 2. Further information about the publicly available datasets can be found in Table 3 and details about the 100-genome dataset can be found in Table 4.

Table 1. Information about the simulated PacBio datasets

Dataset	Read type	Species present	Genome size (Mb)	Coverage	Abundance	Dataset size (GB)	Average read length (kb)
Zymo-1Y2B	PacBio	*S. cerevisiae*	13.163	15x	4.4%	4.2	8.298
		P. aeruginosa	6.792	550x	82.9%		
		L. fermentum	1.905	300x	12.7%		
Zymo-1Y3B	PacBio	*S. cerevisiae*	13.163	15x	3.4%	5.45	8.297
		P. aeruginosa	6.792	550x	64.6%		
		L. fermentum	1.905	300x	9.9%		
		E. faecalis	2.845	450x	22.1%		
Zymo-2Y2B	PacBio	*S. cerevisiae*	13.163	15x	4.2%	4.35	8.299
		C. neoformans	19.325	10x	4.1%		
		P. aeruginosa	6.792	550x	79.5%		
		L. fermentum	1.905	300x	12.2%		
Zymo-2Y3B	PacBio	*S. cerevisiae*	13.163	15x	3.3%	5.65	8.298
		C. neoformans	19.325	10x	3.2%		
		P. aeruginosa	6.792	550x	62.5%		
		L. fermentum	1.905	300x	9.6%		
		E. faecalis	2.845	450x	21.4%		
Zymo-2Y4B	PacBio	*S. cerevisiae*	13.163	15x	2.6%	7.15	8.294
		C. neoformans	19.325	10x	2.5%		
		P. aeruginosa	6.792	550x	49.0%		
		L. fermentum	1.905	300x	7.5%		
		E. faecalis	2.845	450x	16.8%		
		S. aureus	2.730	600x	21.5%		
Sharon	PacBio	*E. faecalis*	3.069	2370x	72.6%	9.8	8.281
		S. aureus	2.913	677x	19.7%		
		P. rhinitidis	2.562	148x	3.8%		
		C. avidum	2.562	136x	3.5%		
		S. epidermidis	2.536	17x	0.4%		
Coral+Symbio	PacBio	*P. lutea*	561.222	20x	47.2%	27.65	8.865
		Cladocopium C15	628.606	20x	52.8%		

Table 2. Information about the simulated ONT datasets

Dataset	Read type	Species present	Genome size (Mb)	Coverage	Abundance	Dataset size (GB)	Average read length (kb)
Zymo-1Y2B-ONT	ONT	*S. cerevisiae*	13.163	15x	4.4%	5.30	8.330
		P. aeruginosa	6.792	550x	82.9%		
		L. fermentum	1.905	300x	12.7%		
Zymo-1Y3B-ONT	ONT	*S. cerevisiae*	13.163	15x	3.4%	6.50	8.334
		P. aeruginosa	6.792	550x	64.6%		
		L. fermentum	1.905	300x	9.9%		
		E. faecalis	2.845	450x	22.1%		
Zymo-2Y2B-ONT	ONT	*S. cerevisiae*	13.163	15x	4.2%	5.50	8.325
		C. neoformans	19.325	10x	4.1%		
		P. aeruginosa	6.792	550x	79.5%		
		L. fermentum	1.905	300x	12.2%		
Zymo-2Y3B-ONT	ONT	*S. cerevisiae*	13.163	15x	3.3%	6.70	8.333
		C. neoformans	19.325	10x	3.2%		
		P. aeruginosa	6.792	550x	62.5%		
		L. fermentum	1.905	300x	9.6%		
		E. faecalis	2.845	450x	21.4%		
Zymo-2Y4B-ONT	ONT	*S. cerevisiae*	13.163	15x	2.6%	8.20	8.329
		C. neoformans	19.325	10x	2.5%		
		P. aeruginosa	6.792	550x	49.0%		
		L. fermentum	1.905	300x	7.5%		
		E. faecalis	2.845	450x	16.8%		
		S. aureus	2.730	600x	21.5%		

3 Evaluation Criteria

The binning result is represented as a $M \times N$ matrix where M refers to the number of bins and N refers to the number of species. In this matrix, the element R_{ij} denotes the number of reads in bin i and that belong to species j. Let T be the total number of reads binned. The precision, recall,

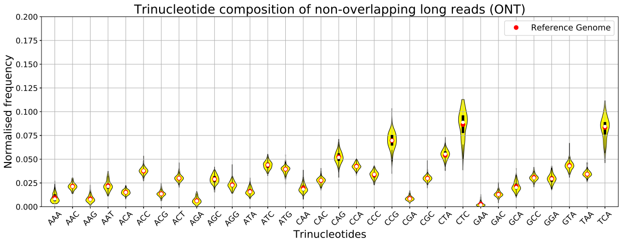

Fig. 1: Trinucleotide composition of 100 non-overlapping Oxford Nanopore (ONT) reads simulated from the reference genome of *P. aeruginosa*. Normalised frequencies are obtained by dividing each trimer occurrence by the total number of trimers observed.

F1 score and Adjusted Rand Index (ARI) are calculated as follows (Girotto *et al.*, 2016; Wang *et al.*, 2012, 2017).

$$Precision(\%) = \frac{\sum_{i=1}^{M} max_j\{R_{ij}\}}{\sum_{i=1}^{M}\sum_{j=1}^{N}\{R_{ij}\}} \times 100 \quad (1)$$

$$Recall(\%) = \frac{\sum_{j=1}^{N} max_i\{R_{ij}\}}{\sum_{i=1}^{M}\sum_{j=1}^{N}\{R_{ij}\} + \textit{Number of unclassified reads}} \times 100 \quad (2)$$

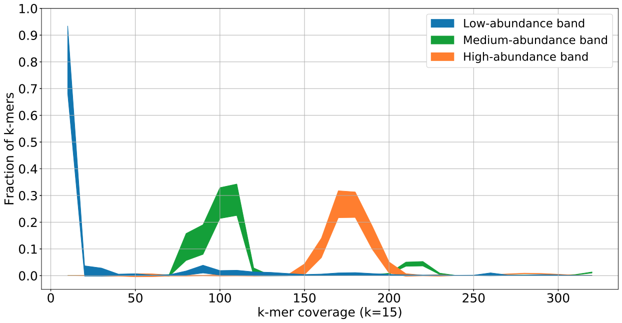

Fig. 2: The k-mer coverage histograms of reads from species of different abundances in the Zymo-1Y3B-ONT dataset

$$\textit{F1 score}(\%) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100 \quad (3)$$

$$ARI(\%) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\binom{R_{ij}}{2} - t_3}{\frac{1}{2}(t_1 + t_2) - t_3} \times 100 \quad (4)$$

$$where\ t_1 = \sum_{i=1}^{M}\binom{\sum_{j=1}^{N} R_{ij}}{2},\ t_2 = \sum_{j=1}^{N}\binom{\sum_{i=1}^{M} R_{ij}}{2},\ and\ t_3 = \frac{t_1 t_2}{\binom{T}{2}}$$

Table 3. Information about the publicly available datasets. †The coverage values for the Zymo-All dataset were obtained from Kolmogorov et al. (2019). *The coverage values for the ASM datasets were obtained from the NCBI SRA taxonomy analysis.

Dataset	Read type	Species present	Genome size (Mb)	Coverage	Abundance	Dataset size (GB)	Average read length (kb)
Zymo-All†	ONT	*P. aeruginosa*	6.792	155x	9.7%	14.24	4.079
		E. coli	4.875	220x	9.9%		
		S. enterica	4.760	227x	10.0%		
		L. fermentum	1.905	528x	9.3%		
		E. faecalis	2.845	464x	12.2%		
		S. aureus	2.730	445x	11.2%		
		L. monocytogenes	2.992	525x	14.5%		
		B. subtilis	4.045	516x	19.3%		
		S. cerevisiae	13.163	17x	2.1%		
		C. neoformans	19.325	10x	1.8%		
ASM-0*	PacBio	*P. aeruginosa*	6.631	5.8x	36.0%	0.10	10.601
		A. pittii	3.917	5.9x	21.6%		
		S. epidermidis	2.535	6.1x	14.5%		
		C. acnes	2.524	6.1x	14.4%		
		S. mitis	2.177	6.6x	13.5%		
ASM-5*	PacBio	*P. aeruginosa*	6.631	5.4x	36.1%	0.10	10.313
		A. pittii	3.917	5.5x	21.7%		
		S. epidermidis	2.535	5.6x	14.3%		
		C. acnes	2.524	5.7x	14.5%		
		S. mitis	2.177	6.1x	13.4%		
ASM-10*	PacBio	*P. aeruginosa*	6.631	5.1x	36.0%	0.10	10.322
		A. pittii	3.917	5.2x	21.7%		
		S. epidermidis	2.535	5.3x	14.3%		
		C. acnes	2.524	5.4x	14.5%		
		S. mitis	2.177	5.8x	13.4%		
ASM-15*	PacBio	*P. aeruginosa*	6.631	4.8x	35.9%	0.10	10.330
		A. pittii	3.917	4.9x	21.7%		
		S. epidermidis	2.535	5.0x	14.3%		
		C. acnes	2.524	5.1x	14.5%		
		S. mitis	2.177	5.5x	13.5%		

Table 4. Information about the 100-genomes dataset. Relative abundance ratios were used according to the simMC+ dataset (Wu et al., 2014)

NCBI Genbank ID	Species present	Relative abundance ratios
256653503	*Acetobacter pasteurianus*	14.5%
330827700	*Aeromonas veronii*	14.5%
398314590	*Amycolatopsis mediterranei*	11.6%
308175814	*Arthrobacter arilaitensis*	7.0%
158421624	*Azorhizobium caulinodans*	4.7%
217957581	*Bacillus cereus*	4.3%
296500838	*Bacillus thuringiensis*	1.2%
42521650	*Bdellovibrio bacteriovorus*	0.6%
119025018	*Bifidobacterium adolescentis*	0.6%
295793053	*Bifidobacterium animalis*	0.6%
343385146	*Brachyspira intermedia*	0.5%
15791399	*Campylobacter jejuni*	0.5%
71082709	*Candidatus Pelagibacter ubique*	0.5%
194246403	*Candidatus Phytoplasma mali*	0.5%
256370581	*Candidatus Sulcia muelleri*	0.5%
297749010	*Chlamydia trachomatis*	0.5%
334694771	*Chlamydophila psittaci*	0.5%
325507407	*Clostridium acetobutylicum*	0.5%
331268188	*Clostridium botulinum*	0.5%
28209834	*Clostridium tetani*	0.5%
125972525	*Clostridium thermocellum*	0.5%
376247367	*Corynebacterium diphtheriae*	0.5%
385806437	*Corynebacterium pseudotuberculosis*	0.5%
334695745	*Corynebacterium ulcerans*	0.5%
284928601	*Cyanobacterium UCYN*	0.5%
307149945	*Cyanothece sp*	0.5%
46562128	*Desulfovibrio vulgaris*	0.5%
58616727	*Ehrlichia ruminantium*	0.5%
378937014	*Enterococcus faecium*	0.5%
336065242	*Erysipelothrix rhusiopathiae*	0.5%
209917191	*Escherichia coli*	0.5%
385805051	*Fervidicoccus fontis*	0.5%
302325342	*Fibrobacter succinogenes*	0.5%
347534971	*Flavobacterium branchiophilum*	0.5%
118496615	*Francisella novicida*	0.5%
156501369	*Francisella tularensis*	0.5%
19703352	*Fusobacterium nucleatum*	0.5%
333392846	*Gardnerella vaginalis*	0.5%
322433659	*Granulicella tundricola*	0.5%
148826757	*Haemophilus influenzae*	0.5%
301154649	*Haemophilus parainfluenzae*	0.5%
170717206	*Haemophilus somnus*	0.5%
12057215	*Halobacterium sp*	0.5%
261854630	*Halothiobacillus neapolitanus*	0.5%
261838873	*Helicobacter pylori*	0.5%
338736863	*Hyphomicrobium sp*	0.5%
385808586	*Ignavibacterium album*	0.5%
375256816	*Klebsiella oxytoca*	0.5%
332290650	*Krokinobacter sp*	0.5%
116332681	*Lactobacillus brevis*	0.5%
327384027	*Lactobacillus casei*	0.5%
104773257	*Lactobacillus delbrueckii*	0.5%
94986445	*Lawsonia intracellularis*	0.5%
296105497	*Legionella pneumophila*	0.5%
330833867	*Metallosphaera cuprina*	0.5%
124484829	*Methanocorpusculum labreanum*	0.5%
19918815	*Methanosarcina acetivorans*	0.5%

Continued to next page

.

NCBI Genbank ID	Species present	Relative abundance ratios
73667559	*Methanosarcina barkeri*	0.5%
239916571	*Micrococcus luteus*	0.5%
356592064	*Mycobacterium bovis*	0.5%
108796981	*Mycobacterium sp*	0.5%
330723203	*Mycoplasma hyorhinis*	0.5%
308388224	*Neisseria meningitidis*	0.5%
300112745	*Nitrosococcus watsonii*	0.5%
325980881	*Nitrosomonas sp*	0.5%
54021964	*Nocardia farcinica*	0.5%
325278757	*Odoribacter splanchnicus*	0.5%
386720569	*Paenibacillus mucilaginosus*	0.5%
261403876	*Paenibacillus sp*	0.5%
54307237	*Photobacterium profundum*	0.5%
126695337	*Prochlorococcus marinus*	0.5%
347537839	*Pseudogulbenkiania sp*	0.5%
313496345	*Pseudomonas putida*	0.5%
116249766	*Rhizobium leguminosarum*	0.5%
111017022	*Rhodococcus jostii*	0.5%
380760311	*Rickettsia prowazekii*	0.5%
378722019	*Rickettsia rickettsii*	0.5%
374318767	*Rickettsia slovaca*	0.5%
99079841	*Ruegeria sp*	0.5%
194447306	*Salmonella enterica*	0.5%
269118642	*Sebaldella termitidis*	0.5%
114045513	*Shewanella sp*	0.5%
30061571	*Shigella flexneri*	0.5%
85057978	*Sodalis glossinidius*	0.5%
311222926	*Staphylococcus aureus*	0.5%
182682970	*Streptococcus pneumoniae*	0.5%
28894912	*Streptococcus pyogenes*	0.5%
354984442	*Streptococcus suis*	0.5%
116626972	*Streptococcus thermophilus*	0.5%
290954631	*Streptomyces scabiei*	0.5%
51891138	*Symbiobacterium thermophilum*	0.5%
320114857	*Thermoanaerobacter brockii*	0.5%
307723218	*Thermoanaerobacter sp*	0.5%
242397997	*Thermococcus sibiricus*	0.5%
239819985	*Variovorax paradoxus*	0.5%
323436265	*Weeksella virosa*	0.5%
225629872	*Wolbachia sp*	0.5%
154243958	*Xanthobacter autotrophicus*	0.5%
162418099	*Yersinia pestis*	0.5%

4 Results of the ONT Read Datasets

To demonstrate how MetaBCC-LR handles Nanopore reads, all the **Zymo** datasets were simulated with DeepSimulator (Li *et al.*, 2018) according to the **Zymo-All** dataset (Nicholls *et al.*, 2019) and binned using MetaBCC-LR. We binned this dataset using MetaBCC-LR and the evaluation results are tabulated in Table 5 in comparison with the results of the corresponding PacBio datasets.

5 Effect of Initial Sample Size

We selected sample sizes 0.5%, 1% and 1.5% of reads from each of the complete datasets to determine the number of bins and build their corresponding statistical profiles. Then, we calculated the precision, recall, F1-score and ARI for the binned sample of reads and the values can be found in Table 6. It can be clearly observed from Table 6 that the increase in sample size has not improved the evaluation scores. Therefore, MetaBCC-LR uses 1% sampling rate by default to perform binning.

6 Memory Usage and Time Complexity of MetaBCC-LR

MetaBCC-LR uses several performance enhancements including multi-threading and in-memory lookup tables to perform computational steps faster. In the **Step 1**, the k-mer coverage histograms are computed using 15-mer counts of the entire dataset. The k-mers are counted using DSK (Rizk *et al.*, 2013) which can operate with multiple threads. All the 15-mer counts are stored in memory as an array of 4^{15} indices holding unsigned 32-bit integers (sufficiently large to store k-mer counts up to

Table 5. Performance comparison of MetaBCC-LR for PacBio and ONT reads.

Dataset	Read type	No. of Bins	Precision	Recall	F1 score	ARI
Zymo-1Y2B	PacBio	3	**99.47%**	**99.47%**	**99.47%**	**98.87%**
	ONT	3	98.99%	98.99%	98.99%	97.63%
Zymo-1Y3B	PacBio	4	**99.27%**	**99.27%**	**99.27%**	**98.57%**
	ONT	4	98.90%	98.90%	98.90%	97.56%
Zymo-2Y2B	PacBio	4	**99.51%**	**99.51%**	**99.51%**	**98.28%**
	ONT	4	99.11%	99.11%	99.11%	97.93%
Zymo-2Y3B	PacBio	5	**99.24%**	**99.24%**	**99.24%**	97.78%
	ONT	5	98.85%	98.85%	98.85%	**97.84%**
Zymo-2Y4B	PacBio	6	**98.46%**	**98.46%**	**98.46%**	**97.21%**
	ONT	6	93.57%	93.57%	93.57%	88.76%

Table 6. Comparison of evaluation metrics for varying sample sizes of the simulated Zymo datasets.

Dataset	Sample size	No. of bins identified	Precision	Recall	F1 score	ARI
Zymo-1Y2B	0.5%	3	99.47%	99.47%	99.47%	98.30%
	1%	3	**99.47%**	**99.47%**	**99.47%**	**98.87%**
	1.5%	3	99.47%	99.47%	99.47%	98.31%
Zymo-1Y3B	0.5%	4	98.16%	98.16%	98.16%	95.46%
	1%	4	**99.27%**	**99.27%**	**99.27%**	**98.57%**
	1.5%	4	99.21%	99.21%	99.21%	97.87%
Zymo-2Y2B	0.5%	4	98.74%	98.74%	98.74%	97.55%
	1%	4	**99.51%**	**99.51%**	**99.51%**	**98.28%**
	1.5%	4	99.31%	99.31%	99.31%	98.29%
Zymo-2Y3B	0.5%	5	98.24%	98.24%	98.24%	97.58%
	1%	5	**99.24%**	**99.24%**	**99.24%**	**97.78%**
	1.5%	5	99.10%	99.10%	99.10%	98.13%
Zymo-2Y4B	0.5%	6	97.82%	97.82%	97.82%	96.28%
	1%	6	**98.46%**	**98.46%**	**98.46%**	**97.21%**
	1.5%	6	98.10%	98.10%	98.10%	96.75%

2^{32}). This enables the O(1) time lookup of 15-mer counts. This requires an initial memory allocation of 4GB which is a reasonable allocation given the performance gain compared to a much slower binary search.

Conversion of reads into 15-mer coverage histograms in **Step 1** and computation of trinucleotide composition profiles in **Step 3** are performed in batches of 100,000 reads with multiple threads (8 by default). This will require roughly 1GB of memory for **Step 1** and **Step 2** (on top of 4GB in the **Step 1** for an average read length of 10,000bp). Raw data is always converted into binary representation of 2 bits per nucleotide.

Steps 2 and **Step 4** of BH-tSNE (Van Der Maaten, 2014) dimension reduction and DB-SCAN (Ester *et al.*, 1996) clustering run on a single thread with $O(Nlog(N))$ and $O(N.d)$ respectively, where N is the number of data points and d is the number of dimensions (Note that $d{=}2$ in these steps). DB-SCAN is performed using multiple threads (8 by default).

Step 5 involves the assignment of all the reads into the bins identified. This is performed in batches of 100,000 reads with 8 threads by default using approximately 384MB of memory. This is because the final classification is performed using the numeric vectors obtained in **Step 1** and **Step 3**. In conclusion, all the steps of MetaBCC-LR are performed under 5GB of peak memory usage.

References

Ester, M. *et al.* (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD'96*, pages 226–231. AAAI Press.

Girotto, S. *et al.* (2016). MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics*, **32**(17), i567–i575.

Kolmogorov, M. *et al.* (2019). metaFlye: scalable long-read metagenome assembly using repeat graphs. *bioRxiv*.

Li, Y. *et al.* (2018). DeepSimulator: a deep simulator for Nanopore sequencing. *Bioinformatics*, **34**(17), 2899–2908.

Nicholls, S. M. *et al.* (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience*, **8**(5). giz043.

Rizk, G. *et al.* (2013). DSK: k-mer counting with very low memory usage. *Bioinformatics*, **29**(5), 652–653.

Van Der Maaten, L. (2014). Accelerating t-SNE Using Tree-based Algorithms. *J. Mach. Learn. Res.*, **15**(1), 3221–3245.

Wang, Y. *et al.* (2012). MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics*, **28**(18), i356–i362.

Wang, Y. *et al.* (2017). Improving contig binning of metagenomic data using d2S oligonucleotide frequency dissimilarity. *BMC Bioinformatics*, **18**(1), 425.

Wu, Y.-W. *et al.* (2014). Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, **2**(1), 26.