3CAC: improving the classification of phages and plasmids from metagenomic assemblies using assembly graphs

Lianrong Pu and Ron Shamir

Blavatnik School of Computer Science, Tel Aviv University
lianrongpu@mail.tau.ac.il, rshamir@tau.ac.il

Abstract. Bacteriophages and plasmids usually coexist with their host bacteria in microbial communities and play important roles in microbial evolution. Accurately identifying sequence contigs as phages, plasmids, and bacterial chromosomes in mixed metagenomic assemblies is critical for further unravelling their functions. Many classification tools have been developed for identifying either phages or plasmids in metagenomic assemblies. However, only two classifiers, PPR-Meta and viralVerify, were proposed to simultaneously identify phages and plasmids in mixed metagenomic assemblies. Due to the very high fraction of chromosome contigs in the assemblies, both tools achieve high precision in the classification of chromosomes but perform poorly in classifying phages and plasmids. Short contigs in these assemblies are often wrongly classified or classified as uncertain.

Here we present 3CAC, a new three-class classifier that improves the precision of phage and plasmid classifications. 3CAC starts with an initial three-class classification generated by existing classifiers and further improves the classification of short contigs and contigs with low confidence classification by using proximity in the assembly graph. Evaluation on simulated metagenomes and on real human gut microbiome samples showed that 3CAC outperformed PPR-Meta and viralVerify in both precision and recall, and increased F1-score by at least 10 percentage points.

Keywords: Metagenome · Three-class Classification · Assembly Graph · Phages · Plasmids.

1 Introduction

The metagenomes of microbial communities are mainly composed of bacterial chromosomes and the associated extrachromosomal mobile genetic elements (eMGEs), such as plasmids and bacteriophages (phages). These eMGEs carry genes related to antibiotic resistance [6, 36, 19], virulence factors [14, 30] and auxiliary metabolic pathways [11, 28, 12]. They can frequently move between species in the microbial community [32, 8] and enable their hosts to rapidly adapt to environmental changes [35, 33]. Despite their important roles in horizontal gene transfer events and in antibiotic resistance, our understanding of these eMGEs is still limited. Part of the difficulty is the challenge of identifying such elements efficiently from mixed metagenomic assemblies [3, 16, 1, 2, 24, 34, 38].

Multiple algorithms have been developed for identifying either phages or plasmids from metagenomic assemblies in recent years. VirSorter and VirSorter2 identify viral metagenomic fragments by searching for reference homologs and testing enrichment of virus-like proteins [29, 10]. These knowledge-based tools have high precision in virus classification but poor ability to identify novel viruses, due to reference database-associated bias. Other tools, such as deepVirFinder [27], Seeker [4], and VIBRANT [12], use machine learning to learn k-mer signatures of viral sequences and perform better on novel virus classification, since they are more loosely linked to annotation databases. cBar is the first tool designed primarily for plasmid identification in metagenomes [40]. More recently, two supervised-learning approaches, PlasFlow [15] and PlasClass [23], were shown to classify plasmid fragments better from metagenomic assemblies. Although both phages and plasmids are commonly found in the metagenomes of microbial communities, all of these tools identify either only phages or only plasmids from metagenomic assemblies.

Currently, only two published tools, PPR-Meta [7] and viralVerify [2], can identify phages and plasmids simultaneously from metagenomic assemblies. However, due to the overwhelming abundance of chromosome fragments in the assemblies (usually $\geq 70\%$), both tools achieve high precision in chromosome classification but very low precision in classification of phages and plasmids [7, 2]. Moreover, classification of short contigs is challenging for all the existing classifiers, as they analyze each contig independently [2, 7, 15, 29, 26]. Here we present 3CAC (3-Class Adjacency based Classifier), an algorithm that employs existing two-class and three-class classifiers to generate an initial three-class classification with high precision, and then improves the classification of short contigs and of contigs classified with lower confidence by taking advantage of classification of their neighbors in the assembly graph. Evaluation on simulated and real metagenome datasets with short and long reads showed that 3CAC improved both precision and recall, and increased F1-score by at least 10 percentage points.

2 Methods

3CAC accepts as input a set of contigs and its associated assembly graph, uses the classification result of existing tools as a starting point, and repeatedly improves

the classification using the assembly graph. Its output is a classification of each contig in the input as phage, plasmid, chromosome, or uncertain. The details of the algorithm are described below.

2.1 Generating the initial classification

3CAC exploits existing two-class and three-class classifiers to generate an initial three-class classification as follows.

(1) Generating a three-class classification. The algorithm runs either viralVerify or PPR-Meta on the set of the input contigs and classifies each contigs as phage, plasmid, chromosome, or uncertain. viralVerify was designed to classify contigs as viral, non-viral or uncertain. Moreover, for non-viral contigs, viralVerify can further classify them as plasmid or non-plasmid using *-p* option. Here, we used *-p* option of viralVerify to classify each of the input contigs as viral, plasmid, chromosome, or uncertain. PPR-Meta calculates three scores representing the probabilities of a contig to be classified as a phage, plasmid, or chromosome. By default, PPR-Meta classifies a contig into the class with the highest score. If a specified score threshold is provided and no score passes the threshold, the sequence will be classified as uncertain. Here, we ran PPR-Meta with a score threshold of 0.7.

(2) Improving plasmid classification. To improve the precision of plasmid classification, PlasClass is run on contigs classified as plasmids in step (1). PlasClass outputs for each contig the probability that it originated from a plasmid. By default, PlasClass classifies a contig as plasmid if it has a probability > 0.5 and as chromosome otherwise. To assure high precision, here we identify contigs with probability ≥ 0.7 as plasmids. Contigs with probability ≤ 0.3 are moved to the chromosome class. The remaining contigs are reclassified as uncertain.

(3) Improving phage classification. Similarly, in order to improve the precision of phage classification, we run deepVirFinder on all contigs classified as phages in step (1). deepVirFinder generates a score and a p-value for each input contig. Contigs with higher scores or lower p-values are more likely to be viral sequences. Here, a contig is kept in the phage class if its p-value ≤ 0.03 and moved to the chromosome class if its p-value > 0.03 and its score ≤ 0.5. The remaining contigs are reclassified as uncertain.

We will denote the algorithm up to this step **Initial(vV)** and **Initial(PM)** if viralVerify or PPR-Meta were used in step (1), respectively.

2.2 Refining the classification using the assembly graph

In genomics and metagenomics, assembly graphs, such as de Bruijn graphs [18, 25] and string graphs [21, 31], are used as the core data structure to combine overlapped reads (or k-mers) into contigs. Nodes in an assembly graph represent contigs and edges represent subsequence overlaps between contigs. Existing classifiers take contigs as input and classify each of them independently based

on its sequence. The overlap information between neighboring contigs in the assembly graphs was ignored by all the existing classifiers. However, recent studies showed that neighboring contigs in an assembly graph are more likely to come from the same taxonomic group [5, 20]. Based on this insight, here we exploit the assembly graph to improve the classification by the following two steps.

(1) Correction of classified contigs. Scan all the classified contigs in the assembly graph in random order. If a classified contig has ≥ 2 classified neighbors and all of them belong to same class, while this contig was classified into a different class, we reason that this contig was wrongly classified and correct its classification to match that of its classified neighbors. This step is repeated until no change was made.

(2) Propagation of the classification to uncertain contigs. Scan all the uncertain contigs in the assembly graph in random order. If an uncertain contig has one or more classified neighbors and all of them belong to same class, we classify this contig into the same class as its classified neighbors. We repeat this step until no uncertain contigs could be classified.

Figure 1 shows the result of applying steps (1) and (2) in a small assembly graph, which is part of the graph generated by assembling simulated long reads (Sim4; see details in the Results section).

We will use the names **3CAC(vV)** and **3CAC(PM)** for the full 3CAC algorithms initialized with viralVerify and PPR-Meta solutions, respectively.

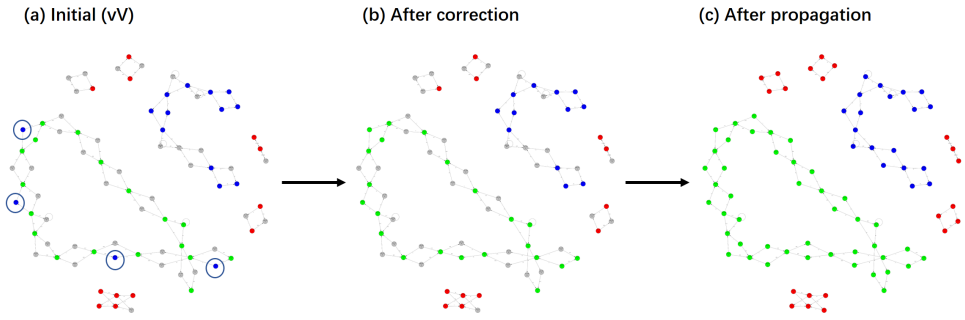


Fig. 1. An example of improving the classification using the assembly graph. Vertices with color red, blue, green, and grey represent contigs classified as phages, plasmids, chromosomes, and uncertain, respectively. (a) The result of Initial(vV). (b) After the correction step. The four contigs encircled in (a) were corrected. (c) After the propagation step.

3 Results

We tested 3CAC on both simulated and real metagenomic assemblies and compared it to PPR-Meta and viralVerify.

3.1 Evaluation criteria

3CAC, viralVerify and PPR-Meta were evaluated based on precision, recall, and F1 score, calculated as follows.

- **Precision:** the fraction of correctly classified contigs among all classified contigs. Note that uncertain contigs were not included in the calculation.
- **Recall:** the fraction of correctly classified contigs among all contigs.
- **F1 score:** the harmonic mean of the precision and recall, which can be calculated as: $F1\ score = (2 * precision * recall)/(precision + recall)$.

Following [23, 7], the precision, recall, and F1 score here were calculated by counting the number of contigs and did not take into account their length. The precision and recall were also calculated separately for phage, plasmid and chromosome classification. For example, the precision of phage classification was calculated as the fraction of correctly classified phage contigs among all contigs classified as phages, and the recall of phage classification was calculated as the fraction of correctly classified phage contigs among all phage contigs.

3.2 Performance on simulated metagenome assemblies

We generated two short-read and two long-read metagenome samples as follows. Sequences of complete bacterial genomes were randomly selected from the NCBI database along with their associated plasmids. The abundance of bacterial genomes was modeled by the log-normal distribution and the copy numbers of plasmids were simulated by the geometric distribution as in [23]. The phage genomes and their abundance profiles were sampled from [26]. Two metagenomic datasets of different complexities were designed. For each of the datasets, 150bp-long short reads were simulated from the genome sequences using InSilicoSeq [9] and assembled by metaSPAdes [22]. Long reads were simulated from the genome sequences using NanoSim [39] and assembled by metaFlye [13]. The error rate of long reads was 9.8% and their average length was 14.9kb. For each assembly, contigs were matched to the reference genomes used in the simulation by minimap2 [17]. Contigs having matches to a reference genome with $\geq 90\%$ mapping identity along $\geq 80\%$ of the contig length were assigned to the class of that reference, and these assignments were used as the gold standard to test the classifiers. Table 1 presents a summary of the simulated metagenome assemblies.

Table 1. Properties of the simulated and the real metagenome datasets and of their assemblies. The number of genome references for the real human gut metagenomes is the number of all complete phage, plasmid and chromosome genomes in NCBI database.

	Read type	Number of reads	# of genome references			# of assembled contigs			Short contigs (< 1kb)
			chromosome	plasmid	phage	chromosome	plasmid	phage	
Sim1	MiSeq	61M	50	193	200	12,494	1,699	696	8,991
Sim2	MiSeq	100M	100	410	500	40,412	5,350	2,926	33,640
Sim3	Nanopore	0.5M	50	193	200	890	166	175	45
Sim4	Nanopore	1M	100	410	500	2,491	395	413	152
Gut microbiome	HiSeq	53.8M	19,053	20,838	13,903	130,252	943	383	110,128
Gut microbiome	Pacbio	14.7M	19,053	20,838	13,903	4,671	64	8	723

Figure 2 shows the performance of PPR-Meta, viralVerify and the first phase of 3CAC on these simulated metagenome assemblies. Both PPR-Meta and viralVerify had high precision in chromosome classification, but their precision in phage and plasmid classification was usually low. Further analysis revealed that both of the algorithms distinguished well between phages and plasmids. Their low precision in phage and plasmid classification was due to contamination from chromosome contigs (Supplementary Table A.1). Utilizing two-class classifiers, PlasClass and deepVirFinder, the first phase of 3CAC improved markedly the precision in phage and plasmid classification, while it decreased a little bit the precision in chromosome classification (Figure 2, Supplementary Table A.2). In contrast, recall decreased in phage and plasmid classification, but increased in chromosome classification (Supplementary Figure B.1).

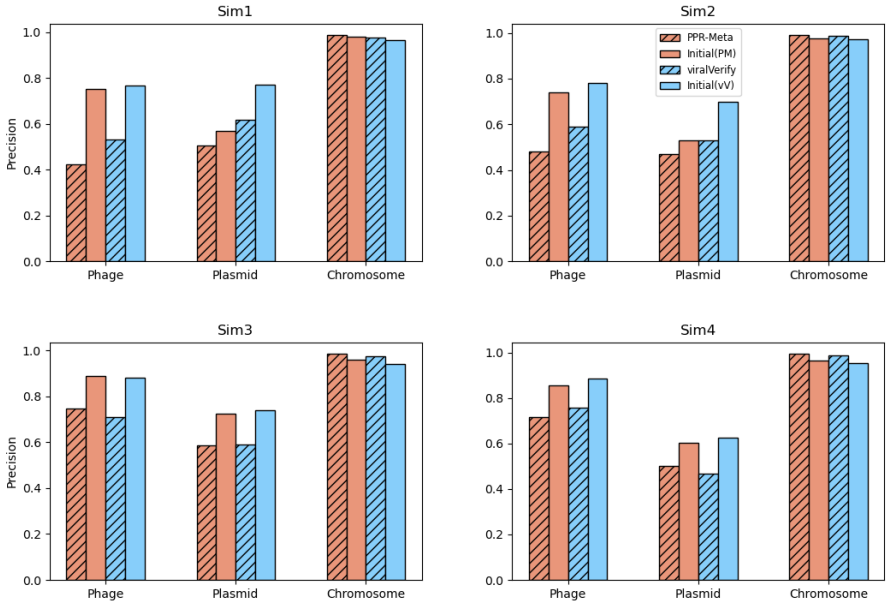


Fig. 2. Precision of the initial classification of 3CAC compared to PPR-Meta and viralVerify. See supplementary Figure B.1 for recall.

Figure 3 shows the results of initial phase of 3CAC on the short-read simulated metagenome assemblies for different contig lengths. Short contigs tended to have lower recall in the initial classification of 3CAC, while precision was not sensitive to the contig length. When the initial classification of 3CAC was generated based on PPR-Meta solution, recall decreased sharply for contigs with length $<$ 1kb. When viralVerify solution was used, recall was even lower for contig shorter

than 1kb and improvement with size was roughly linear. We reasoned that these classifiers classified each of the input contigs independently, and so short contigs could not be classified reliably. However, Table 1 shows that more than half of the contigs assembled from short reads are shorter than 1kb. To assist in the classification of these short contigs, 3CAC was designed to take advantage of the longer contigs with confident classification and that are neighbors of these short contigs in the assembly graph. Figure 3 shows that 3CAC significantly increased recall for all contigs with almost no loss of precision. Remarkably, the recall for contigs shorter than 1kb increased from < 0.2 to ≥ 0.8. For contigs assembled from long reads, 3CAC not only improved the recall substantially but also slightly improved the precision (Figure 4).

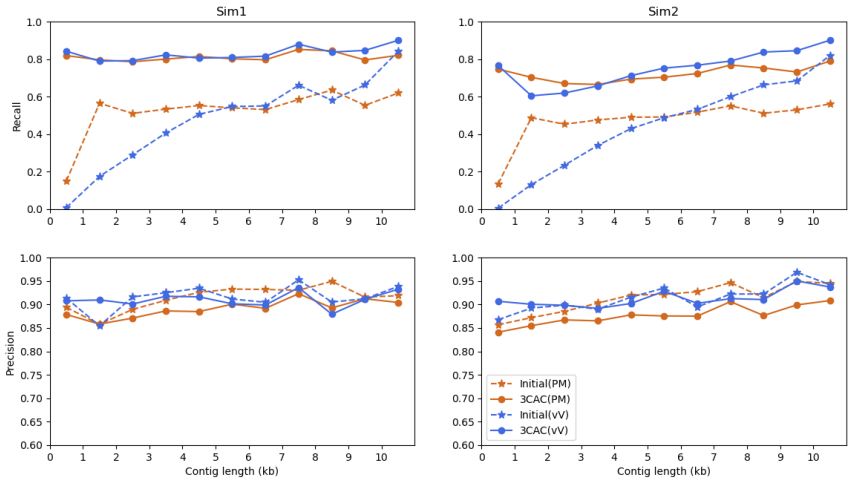


Fig. 3. Performance on contigs assembled from simulated short reads. Results are shown for contigs of lengths < 1 kb, 1-2 kb, ...,9-10 kb, ≥ 10kb .

The analysis above shows that the two phases of 3CAC algorithm improved the precision and recall for the three-class classification. Evaluation of PPR-Meta, viralVerify and 3CAC on these simulated metagenome assemblies showed that 3CAC performed the best in all the assemblies (Figure 5). 3CAC outperformed PPR-Meta and viralVerify in both precision and recall. For contigs assembled from short reads (Sim1 and Sim2), the recall and F1 scores of viralVerify were more than doubled by 3CAC. We also calculated the precision, recall and F1 scores for phage, plasmid, and chromosome classification separately (Supplementary Table A.3). 3CAC(vV) had the best F1 scores on all the datasets and the highest precision in classification of phages and plasmids. Note that PPR-Meta here was run with default setting. Running PPR-Meta with 0.7 score threshold (as done in Initial(PM)) resulted in higher precision but lower recall and lower F1 score. Supplementary Table A.4 shows that 3CAC also outperformed PPR-Meta with 0.7 score threshold.

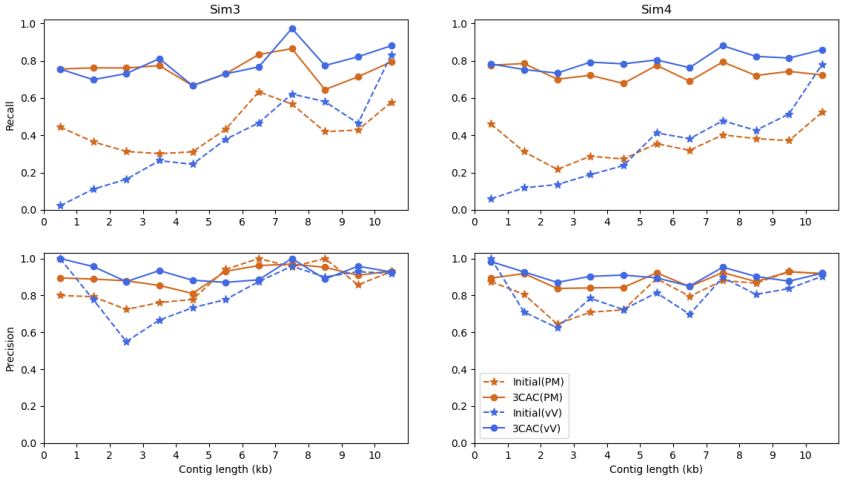


Fig. 4. Performance on contigs assembled from simulated long reads. Results are shown for contigs of lengths < 1 kb, 1-2 kb, ...,9-10 kb, $\geq$ 10kb .

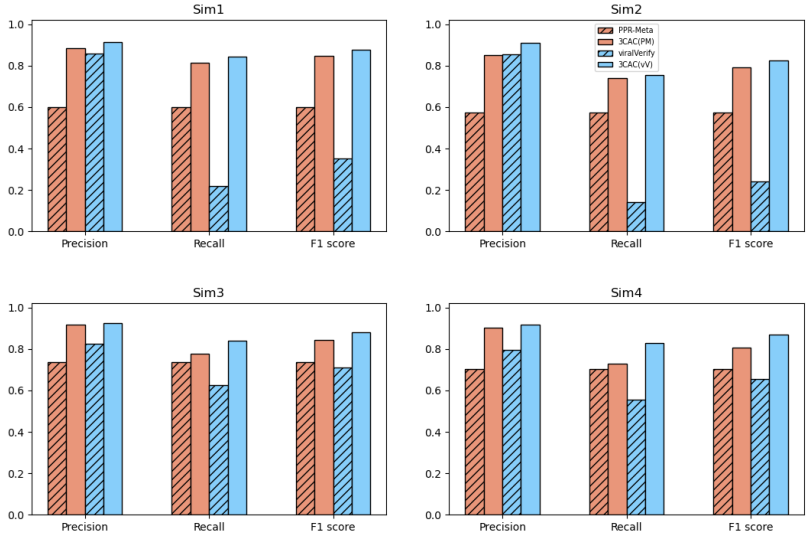


Fig. 5. Performance of three-class classifiers on the simulated metagenome assemblies.

3.3 Performance on human gut microbiome samples

Five publicly available human gut microbiome samples with short-read sequencing datasets (NCBI accession numbers: ERR12976697, ERR1297651, ERR1297751, ERR1297845, ERR1297770) were selected and assembled together using metaSPAdes [22]. Another set of five human gut microbiome samples with long-read sequencing datasets (NCBI accessions: SRX2529348, SRX2529347, SRX2529346, SRX2529341, SRX2529340) were selected from [34] and assembled together using metaFlye [13]. To identify the class of contigs in the real metagenome assemblies, we downloaded all complete phage, plasmid and chromosome genomes from NCBI database and mapped contigs to all the reference genomes using minimap2 [17]. A contig was considered matched to a reference sequence if it had $\geq$ 80% mapping identity along $\geq$ 80% of the contig length. Contigs that matched to reference genomes of two or more classes were excluded to avoid ambiguity. Overall, 131,578 out of 469,022 contigs in the short-read assembly and 4,743 out of 12,541 contigs in the long-read assembly had matches to a single class and were used as the gold standard to test the classifiers. Table 1 summarizes the properties of the datasets and the assemblies.

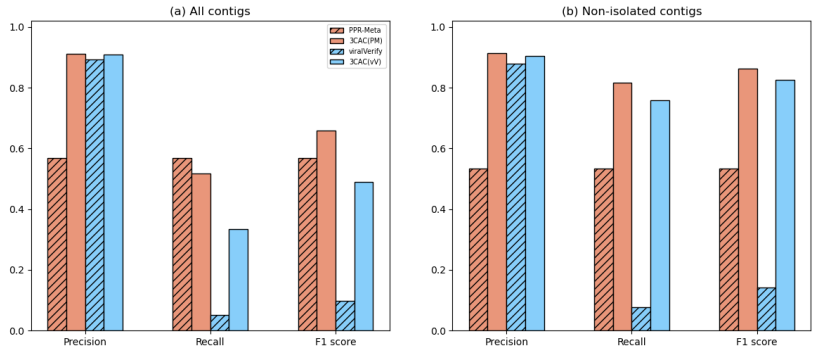


Fig. 6. Performance of three-class classifiers on contigs assembled from short-read sequencing of human gut microbiome samples. (a) performance on all contigs; (b) performance on non-isolated contigs in the assembly graph.

Figures 6(a) and 7(a) show the results of PPR-Meta, viralVerify and 3CAC on the short-read and long-read assemblies, respectively. On the long-read assembly, 3CAC(vV) and 3CAC(PM) had comparable performance. 3CAC was best in precision, recall and F1 score (Figure 7).

Interestingly, on the short-read assembly, 3CAC(PM) and PPR-Meta had higher F1 score than 3CAC(vV) (Figure 6 (a)). Further analysis revealed that this was due to a large number of isolated contigs in the short-read assembly graph. The second phase of 3CAC was only performed on contigs that have

neighbours in the assembly graph. However, 59% of the contigs assembled from short reads were isolated and had no neighbors in the assembly graph, while the fraction on the long-read assembly was only 21%. Figures 6 (b) and 7 (b) show the results on the non-isolated contigs in the assembly graph. For both long read and short read assemblies, 3CAC(PM) and 3CAC(vV) had comparable performance and outperformed PPR-meta and viralVerify in precision, recall, and F1 score.

Supplementary Figure B.2 shows the precision, recall and F1 score separately for phage, plasmid and chromosome classification in both short-read and long-read assemblies. In classification of phages and plasmids, PPR-Meta had the highest recall, but its precision was as low as 0.02. Compared to PPR-Meta, 3CAC had higher precision in classification of phages and plasmids, at the cost of lower recall and tended to have better F1 scores. In chromosome classification, 3CAC performed the best in the long-read assembly while PPR-Meta performed slightly better in the short-read assembly.

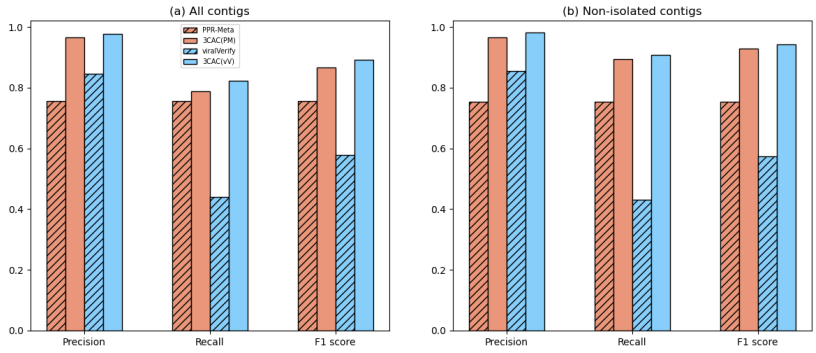


Fig. 7. Performance of three-class classifiers on contigs assembled from long-read sequencing of human gut microbiome samples. (a) performance on all contigs; (b) performance on non-isolated contigs in the assembly graph.

3.4 Software and Resource usage

3CAC uses classifications generated by existing classifiers, and so the running time of its first phase depended on the classifiers used. On our datasets, PPR-Meta, PlasClass and deepVirFinder were fast and required less than an hour. viralVerify took up to 4-5 hours for the real metagenome assemblies with 8 threads. The second phase of 3CAC is fast and took less than 30 minutes in a single thread for all the datasets we tested. Performance was measured on a 44-core, 2.2 GHz server with 792 GB of RAM. 3CAC will be freely available under Shamir-Lab on Github soon.

4 Discussion and Conclusion

Classification of phages and plasmids from mixed metagenome assemblies is important for further unravelling and understanding the functions of these mobile genetic elements in microbiome communities. Many two-class classifiers has been developed in recent years to identify either phages or plasmids from metagenome assemblies. A naive way to identify phages and plasmids simultaneously from mixed metagenome assemblies is by using phage classifiers and plasmid classifiers to identify phages and plasmids respectively, and then combining the classification result. However, this is impractical since phage sequences are often arbitrarily classified as plasmids or chromosomes by plasmid classifiers, and so are the plasmid sequences in phage classifiers. In this work, we first exploit three-class classifiers to accurately separate plasmids from phages and then utilize two-class classfiers to further improve the precision of phage and plasmid classification. The key improvement by 3CAC is obtained by utilizing the structure of the assembly graph to assist the classification of short and uncertain contigs. This leads to significant improvement of the recall and almost no loss of the precision.

Evaluation the performance of classifiers on real metagenome assemblies remains challenging due to the lack of gold standard. By mapping contigs to all the available reference genomes, we are able to identify the class of a fraction of the contigs. However, as shown in previous studies [7], some plasmid genomes are quite similar to their host bacterial chromosomes. Thus, many contigs from metagenome assemblies have matches to both plasmid and chromosome reference genomes, and it is hard to identify their classes. Additionally, many contigs with no matches to the reference database may represent novel species, but they were excluded from our evaluation. Keeping in mind these shortcomings of the gold standard for real metagenome assemblies, 3CAC outperformed existing three-class classifiers substantially.

3CAC is initialized with solutions of PPR-Meta or viralVerify. Their overall performance was comparable, with initialization with PPR-Meta doing slightly better in the short read real data, and initialization with viralVerify slightly better on long read real data and in all simulations. PPR-Meta could be run with different score thresholds, and a higher score threshold results in higher precision and lower recall. In our experiments, we tried the score thresholds 0.7 and 0.8, and the difference in the results was minor.

3CAC has some limitations. The propagation step of 3CAC can greatly improve the recall, but it can only be performed on non-isolated contigs in the assembly graph. The recall of isolated contigs is still limited by the performance of existing classifiers. 3CAC also relies on current 2-class and 3-class classifiers. In the future, we plan to extend 3CAC to a stand-alone classification tool without relying on existing classifiers. Currently, 3CAC scans contigs in the assembly graph in random order, in both the correction and the propagation steps. That order may affect the results, and a more judicious order may improve the classification. Finally, there is room for extending 3CAC to a four-class algorithm that would be able to classify also eukaryotic contigs in metagenome assemblies [37].

Acknowledgments

This study was supported in part by grant 2016694 from the United State - Israel Binational Science Foundation (BSF), Jerusalem, Israel and the United States National Science Foundation (NSF). L.P. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. L.P. was also supported in part by postdoctoral fellowships from the Planning and Budgeting Committee (PBC) of the Council for Higher Education (CHE) in Israel.

References

1. Antipov, D., Raiko, M., Lapidus, A., Pevzner, P.A.: Plasmid detection and assembly in genomic and metagenomic data sets. Genome Research **29**(6), 961–968 (2019)
2. Antipov, D., Raiko, M., Lapidus, A., Pevzner, P.A.: Metaviral SPAdes: assembly of viruses from metagenomic data. Bioinformatics **36**(14), 4126–4129 (2020)
3. Arredondo-Alonso, S., Willems, R.J., Van Schaik, W., Schürch, A.C.: On the (im) possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microbial Genomics **3**(10) (2017)
4. Auslander, N., Gussow, A.B., Benler, S., Wolf, Y.I., Koonin, E.V.: Seeker: alignment-free identification of bacteriophage genomes by deep learning. Nucleic Acids Research **48**(21), e121–e121 (2020)
5. Barnum, T.P., Figueroa, I.A., Carlström, C.I., Lucas, L.N., Engelbrektson, A.L., Coates, J.D.: Genome-resolved metagenomics identifies genetic mobility, metabolic interactions, and unexpected diversity in perchlorate-reducing communities. The ISME Journal **12**(6), 1568–1581 (2018)
6. Calero-Cáceres, W., Ye, M., Balcázar, J.L.: Bacteriophages as environmental reservoirs of antibiotic resistance. Trends in Microbiology **27**(7), 570–577 (2019)
7. Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., Zhu, H.: PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. GigaScience **8**(6), giz066 (2019)
8. Frost, L.S., Leplae, R., Summers, A.O., Toussaint, A.: Mobile genetic elements: the agents of open source evolution. Nature Reviews Microbiology **3**(9), 722–732 (2005)
9. Gourlé, H., Karlsson-Lindsjö, O., Hayer, J., Bongcam-Rudloff, E.: Simulating illumina metagenomic data with insilicoseq. Bioinformatics **35**(3), 521–522 (2019)
10. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., et al.: Virsorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome **9**(1), 1–13 (2021)
11. Hurwitz, B.L., U'Ren, J.M.: Viral metabolic reprogramming in marine ecosystems. Current Opinion in Microbiology **31**, 161–168 (2016)
12. Kieft, K., Zhou, Z., Anantharaman, K.: Vibrant: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome **8**(1), 1–23 (2020)
13. Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P., et al.: metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature Methods **17**(11), 1103–1110 (2020)

14. Kraushaar, B., Hammerl, J., Kienol, M., Heinig, M., Sperling, N., Dinh Thanh, M., Reetz, J., Jackel, C., Fetsch, A., Hertwig, S.: Acquisition of virulence factors in livestock-associated mrsa: lysogenic conversion of cc398 strains by virulence gene-containing phages. Sci Rep 7: 2004 (2017)
15. Krawczyk, P.S., Lipinski, L., Dziembowski, A.: Plasflow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Research **46**(6), e35–e35 (2018)
16. Krishnamurthy, S.R., Wang, D.: Origins and challenges of viral dark matter. Virus Research **239**, 136–142 (2017)
17. Li, H.: Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**(18), 3094–3100 (2018)
18. Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M., Pevzner, P.A.: Assembly of long error-prone reads using de Bruijn graphs. Proceedings of the National Academy of Sciences **113**(52), E8396–E8405 (2016)
19. Lopatkin, A., Meredith, H., Srimani, J., Pfeiffer, C., Durrett, R., You, L.: Persistence and reversal of plasmid-mediated antibiotic resistance. Nature Communications 8: 1689 (2017)
20. Mallawaarachchi, V., Wickramarachchi, A., Lin, Y.: Graphbin: refined binning of metagenomic contigs using assembly graphs. Bioinformatics **36**(11), 3307–3313 (2020)
21. Myers, E.W.: The fragment assembly string graph. Bioinformatics **21**(suppl_2), ii79–ii85 (2005)
22. Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.: metaSPAdes: a new versatile de novo metagenomics assembler. arXiv preprint arXiv:1604.03071 (2016)
23. Pellow, D., Mizrahi, I., Shamir, R.: Plasclass improves plasmid sequence classification. PLoS Computational Biology **16**(4), e1007781 (2020)
24. Pellow, D., Zorea, A., Probst, M., Furman, O., Segal, A., Mizrahi, I., Shamir, R.: Scapp: An algorithm for improved plasmid assembly in metagenomes. Microbiome **9**(1), 1–12 (2021)
25. Pevzner, P.A., Tang, H., Waterman, M.S.: An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences **98**(17), 9748–9753 (2001)
26. Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F.: Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome **5**(1), 1–20 (2017)
27. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R., Sun, F.: Identifying viruses from metagenomic data using deep learning. Quantitative Biology pp. 1–14 (2020)
28. Rosenwasser, S., Ziv, C., Van Creveld, S.G., Vardi, A.: Virocell metabolism: metabolic innovations during host–virus interactions in the ocean. Trends in Microbiology **24**(10), 821–832 (2016)
29. Roux, S., Enault, F., Hurwitz, B.L., Sullivan, M.B.: Virsorter: mining viral signal from microbial genomic data. PeerJ **3**, e985 (2015)
30. Sarowska, J., Futoma-Koloch, B., Jama-Kmiecik, A., Frej-Madrzak, M., Ksiazczyk, M., Bugla-Ploskonska, G., Choroszy-Krol, I.: Virulence factors, prevalence and potential transmission of extraintestinal pathogenic escherichia coli isolated from different sources: recent reports. Gut Pathogens **11**(1), 1–16 (2019)
31. Simpson, J.T., Durbin, R.: Efficient de novo assembly of large genomes using compressed data structures. Genome Research **22**(3), 549–556 (2012)

32. Sitaraman, R.: Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. Microbiome **6**(1), 1–14 (2018)
33. Smalla, K., Jechalke, S., Top, E.M.: Plasmid detection, characterization, and ecology. Microbiology Spectrum **3**(1), 3–1 (2015)
34. Suzuki, Y., Nishijima, S., Furuta, Y., Yoshimura, J., Suda, W., Oshima, K., Hattori, M., Morishita, S.: Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. Microbiome **7**(1), 1–16 (2019)
35. Thomas, C.M., Nielsen, K.M.: Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature Reviews Microbiology **3**(9), 711–721 (2005)
36. Wein, T., Hülter, N., Mizrahi, I., Dagan, T.: Emergence of plasmid stability under non-selective conditions maintains antibiotic resistance. Nature Communications 10: 2595 (2019)
37. West, P.T., Probst, A.J., Grigoriev, I.V., Thomas, B.C., Banfield, J.F.: Genome-reconstruction for eukaryotes from complex natural microbial communities. Genome Research **28**(4), 569–580 (2018)
38. Yahara, K., Suzuki, M., Hirabayashi, A., Suda, W., Hattori, M., Suzuki, Y., Okazaki, Y.: Long-read metagenomics using promethion uncovers oral bacteriophages and their interaction with host bacteria. Nature Communications **12**(1), 1–12 (2021)
39. Yang, C., Chu, J., Warren, R.L., Birol, I.: Nanosim: nanopore sequence read simulator based on statistical characterization. GigaScience **6**(4), gix010 (2017)
40. Zhou, F., Xu, Y.: cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics **26**(16), 2051–2052 (2010)

Supplementary Material

A Supplementary tables

Table A.1. Performance of viralVerify and PPR-Meta on simulated metagenomic assemblies. PPR-Meta was run with a score threshold ≥ 0.7 to assure high precision.

	True Category	# contigs	viralVerify classification				PPR-Meta classification			
			phage	plasmid	chromosome	uncertain	phage	plasmid	chromosome	uncertain
Sim1	phage	696	279	6	1	410	572	1	1	122
	plasmid	1,699	23	380	58	1,238	80	452	43	1,124
	chromosome	12,494	224	231	2,624	9,415	696	442	3,749	7,607
Sim2	phage	2,926	737	7	3	2,179	1,919	14	2	991
	plasmid	5,350	53	652	60	4,585	158	1,351	71	3,770
	chromosome	40,412	462	572	5,448	33,930	1,926	1,520	9,568	27,398
Sim3	phage	175	163	2	0	10	164	0	0	11
	plasmid	166	9	124	12	21	1	88	5	72
	chromosome	890	58	84	485	263	55	62	366	407
Sim4	phage	413	376	4	0	33	362	1	0	50
	plasmid	395	6	301	16	72	2	197	4	192
	chromosome	2,491	114	337	1,161	879	142	196	889	1,264

Table A.2. Performance of Initial(vV) and Initial(PM) on simulated metagenomic assemblies.

	True Category	# contigs	Initial(vV)				Initial(PM)			
			phage	plasmid	chromosome	uncertain	phage	plasmid	chromosome	uncertain
Sim1	phage	696	249	3	8	436	442	0	8	246
	plasmid	1,699	7	310	88	1,294	10	372	75	1,242
	chromosome	12,494	69	90	2,797	9,538	137	284	3,936	8,137
Sim2	phage	2,926	653	4	35	2,234	1,136	9	45	1,736
	plasmid	5,350	18	499	132	4,701	33	964	213	4,140
	chromosome	40,412	165	213	5,803	34,231	363	842	10,092	29,115
Sim3	phage	175	139	1	11	24	145	0	9	21
	plasmid	166	3	108	22	33	0	82	9	75
	chromosome	890	16	37	540	297	18	31	407	434
Sim4	phage	413	321	2	34	71	318	0	23	87
	plasmid	395	2	262	36	103	1	175	14	213
	chromosome	2,491	38	153	1,280	1012	50	114	970	1,349

Table A.3. Classification of phages, plasmids and chromosomes on simulated metagenome assemblies. PPR-Meta and PPR-Meta(0.7) represent running PPR-Meta on default setting and with a score threshold of 0.7, respectively.

	Tool	Phage			Plasmid			Chromosome		
		precision(%)	recall(%)	F1 score(%)	precision(%)	recall(%)	F1 score(%)	precision(%)	recall(%)	F1 score(%)
Sim1	viralVerify	53.04	40.09	45.66	61.59	22.37	32.82	97.8	21.0	34.58
	3CAC(vV)	63.06	92.24	74.91	71.51	63.98	67.54	96.42	86.67	91.28
	PPR-Meta(0.7)	42.43	82.18	55.97	50.5	26.6	34.85	98.84	30.01	46.04
	3CAC(PM)	63.58	92.82	75.47	56.68	65.39	60.73	96.35	83.1	89.24
	PPR-Meta	19.06	94.54	31.72	25.97	57.74	35.83	95.31	58.43	72.45
Sim2	viralVerify	58.87	25.19	35.28	52.97	12.19	19.81	98.86	13.48	23.73
	3CAC(vV)	74.34	92.79	82.55	68.46	49.78	57.64	95.44	77.72	85.68
	PPR-Meta (0.7)	47.94	65.58	55.39	46.83	25.25	32.81	99.24	23.68	38.23
	3CAC(PM)	66.55	92.21	77.31	46.81	52.93	49.68	94.82	75.45	84.03
	PPR-Meta	21.83	90.46	35.17	24.31	61.7	34.88	95.86	54.53	69.52
Sim3	viralVerify	70.87	93.14	80.49	59.05	74.7	65.96	97.59	54.49	69.94
	3CAC(vV)	91.08	81.71	86.14	80.0	74.7	77.26	95.39	86.07	90.49
	PPR-Meta (0.7)	74.55	93.71	83.04	58.67	53.01	55.70	98.65	41.12	58.05
	3CAC(PM)	87.57	84.57	86.05	74.83	66.27	70.29	96.42	78.76	86.70
	PPR-Meta	60.63	99.43	75.33	42.04	84.34	56.11	97.38	66.85	79.28
Sim4	viralVerify	75.81	91.04	82.73	46.88	76.2	58.05	98.64	46.61	63.30
	3CAC(vV)	90.32	81.36	85.61	68.76	77.47	72.86	96.49	83.98	89.80
	PPR-Meta (0.7)	71.54	87.65	78.78	50.0	49.87	49.94	99.55	35.69	52.54
	3CAC(PM)	81.59	79.42	80.49	61.47	52.91	56.87	97.04	75.07	84.65
	PPR-Meta	60.3	98.55	74.82	33.17	84.05	47.57	97.47	63.51	76.91

Table A.4. Performance of three-class classifiers on simulated metagenomic assemblies. PPR-Meta and PPR-Meta(0.7) represent running PPR-Meta on default setting and with a score threshold of 0.7, respectively.

Dataset	Evaluation Criteria	**viralVerify**	**3CAC(vV)**	**PPR-Meta(0.7)**	**3CAC(PM)**	**PPR-Meta**
Sim1	Precision	85.81%	**91.2%**	79.08%	88.28%	60.04%
	Recall	22.05%	**84.34%**	32.06%	81.54%	60.04%
	F1 score	35.08%	**87.64%**	45.62%	84.77%	60.04%
Sim2	Precision	85.53%	**90.94%**	77.67%	85.23%	57.47%
	Recall	14.04%	**75.56%**	26.37%	73.98%	57.47%
	F1 score	24.12%	**82.54%**	39.37%	79.21%	57.47%
Sim3	Precision	82.39%	**92.65%**	83.4%	91.95%	73.84%
	Recall	62.71%	**83.92%**	50.2%	77.9%	73.84%
	F1 score	71.22%	**88.06%**	62.68%	84.34%	73.84%
Sim4	Precision	79.4%	**91.59%**	80.76%	90.18%	70.35%
	Recall	55.71%	**82.87%**	43.89%	72.96%	70.35%
	F1 score	65.48%	**87.01%**	56.87%	80.66%	70.35%

B Supplementary figures

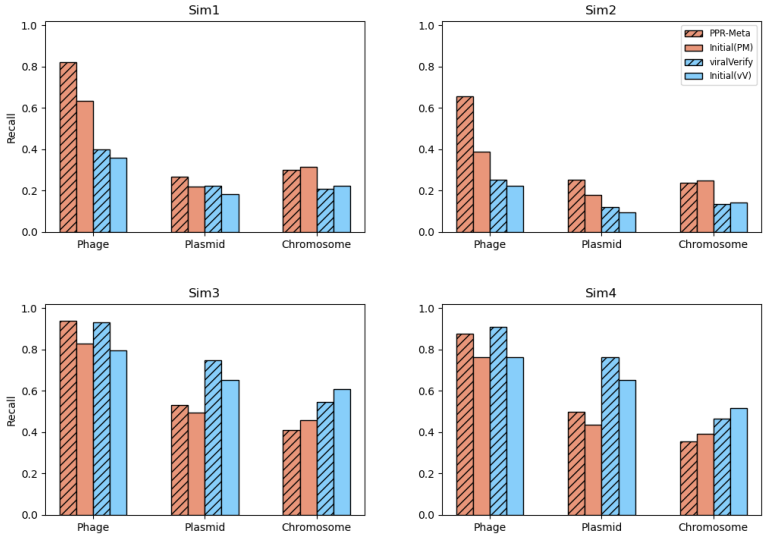


Fig. B.1. Recall of the initial classification of 3CAC compared to PPR-Meta and viralVerify.

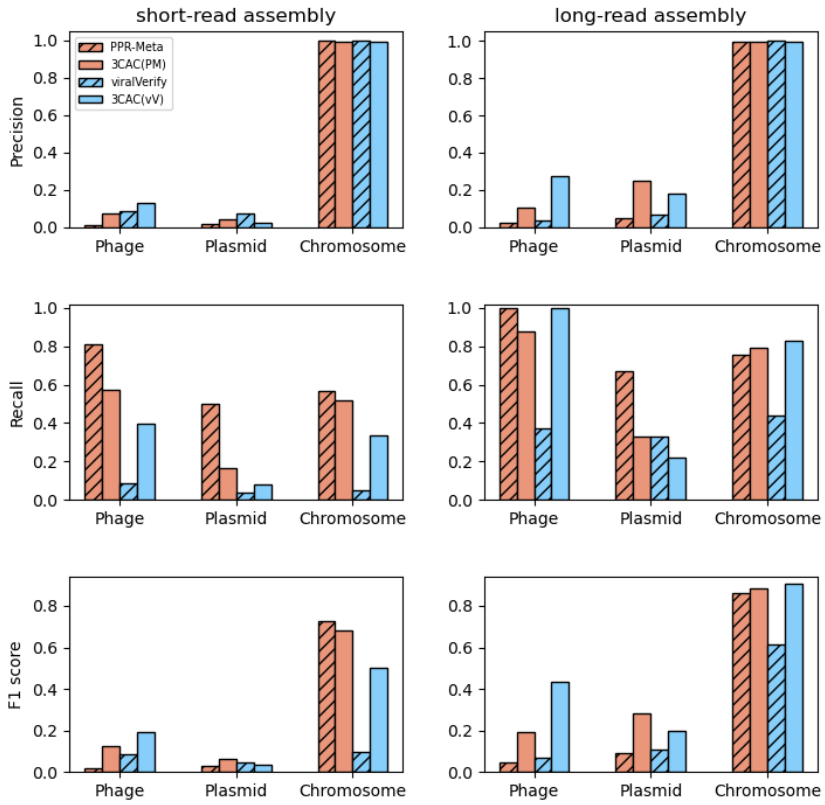
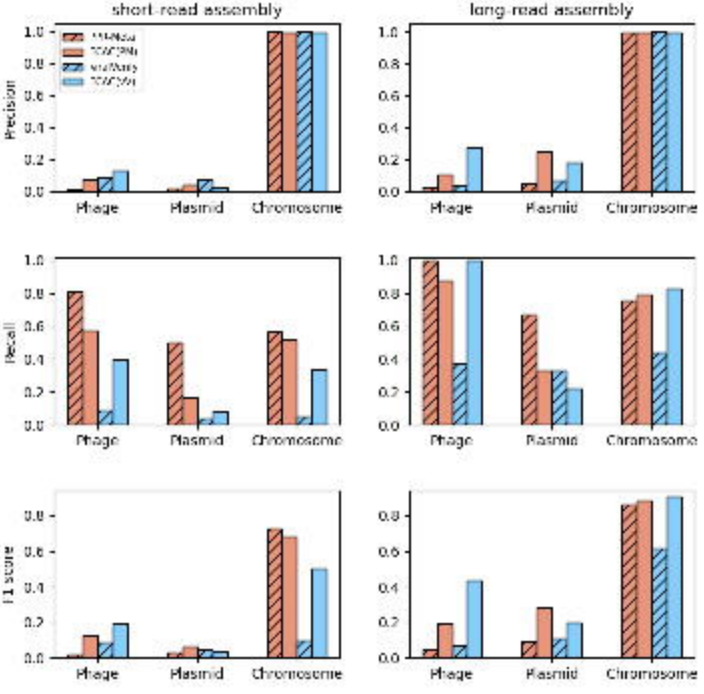
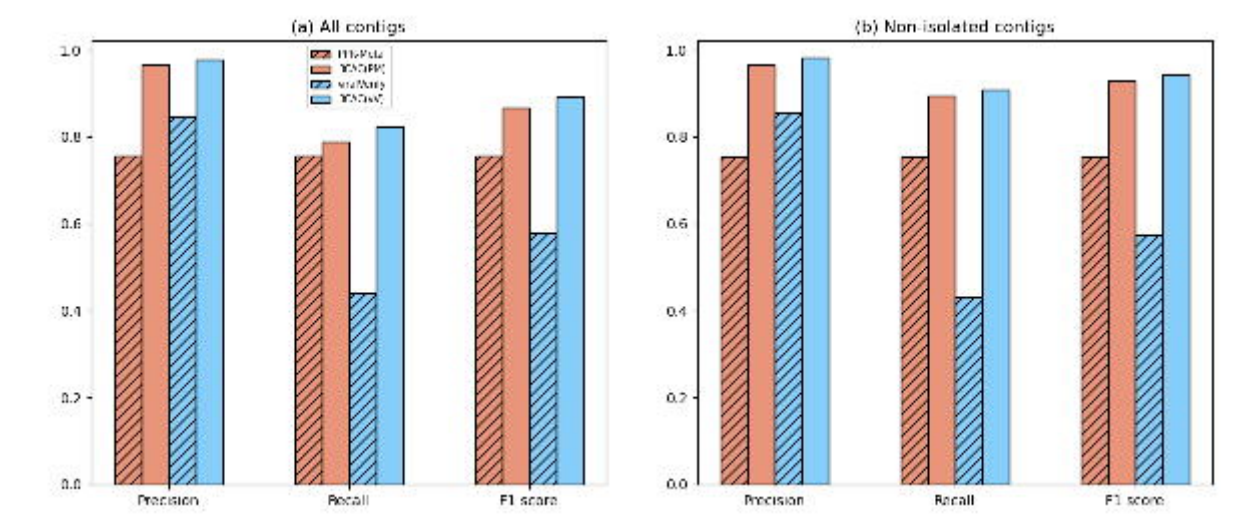
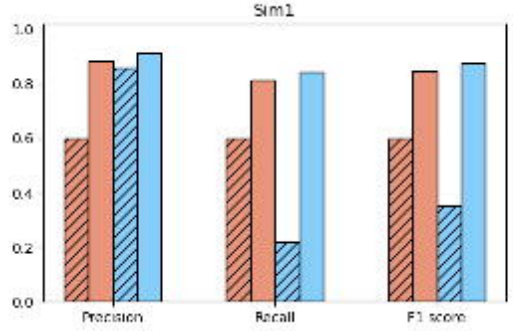
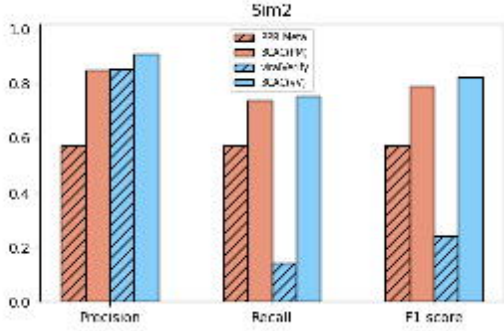


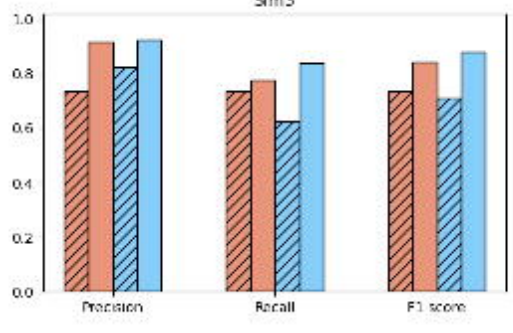
Fig. B.2. Performance of three-class classifiers on real human gut samples with short-read and long-read assemblies.

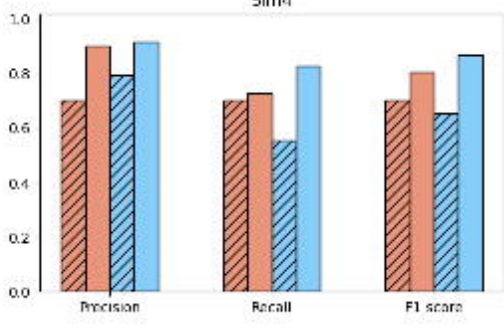

short-read assembly
long-read assembly
Precision
Recall
F1 score
Phage
Plasmid
Chromosome

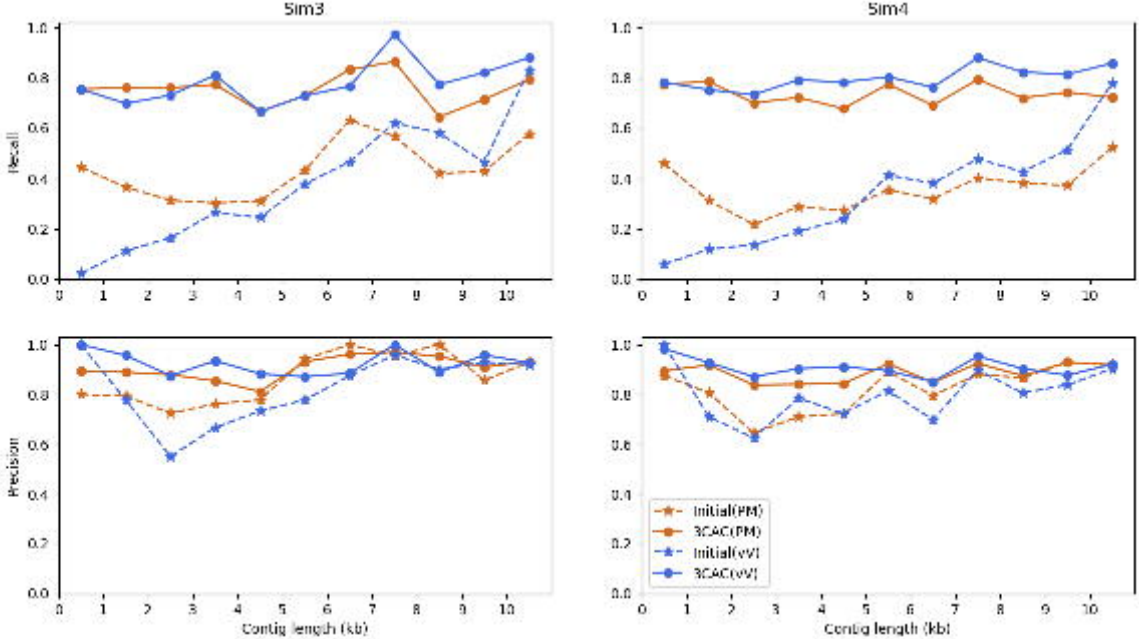

(a) All contigs
(b) Non-isolated contigs
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1 score

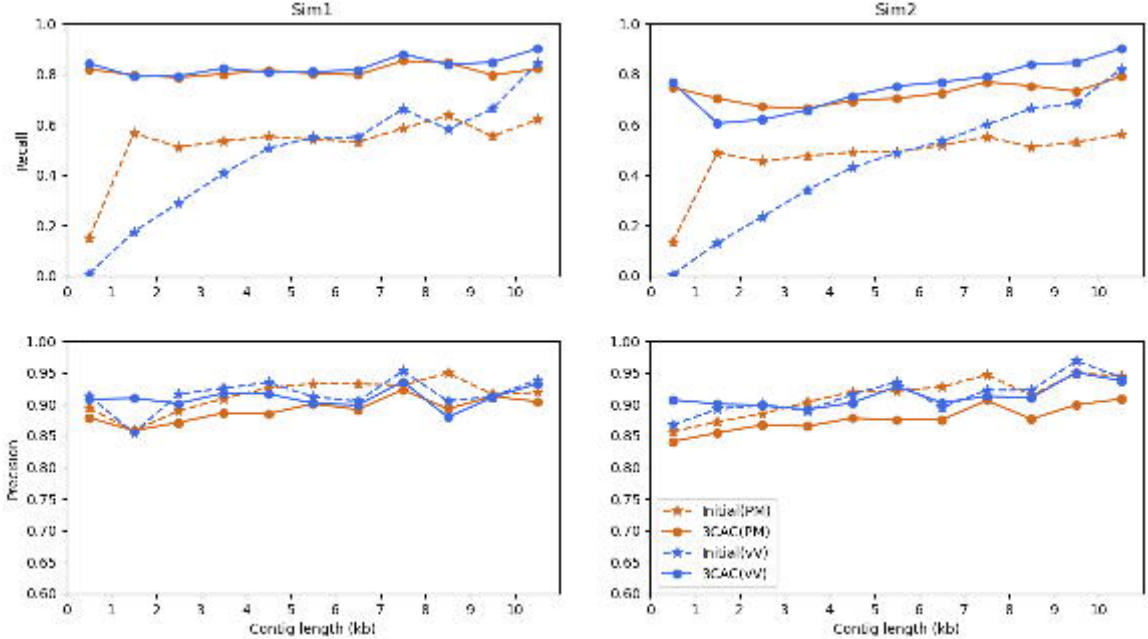

Sim1
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1 score

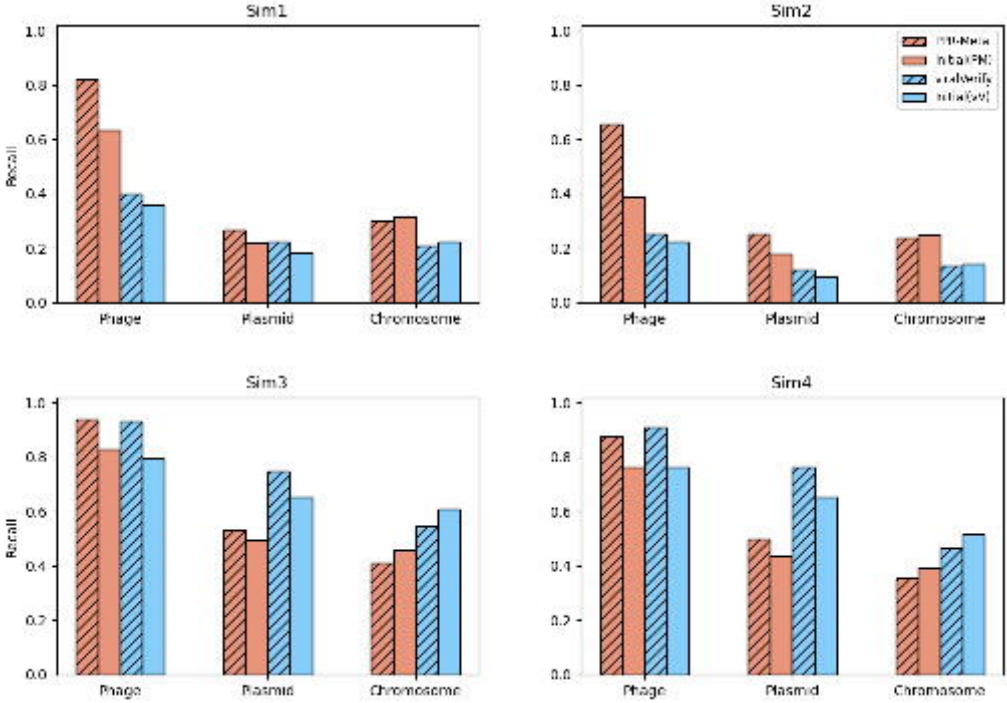

Sim2
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1 score

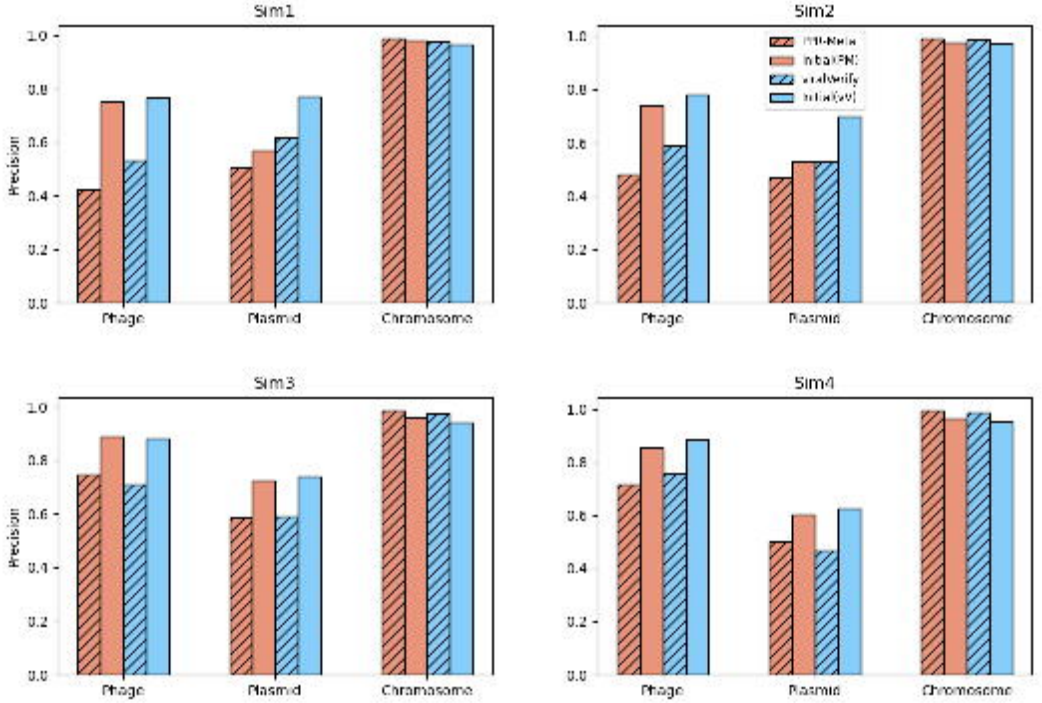

Sim3
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1 score

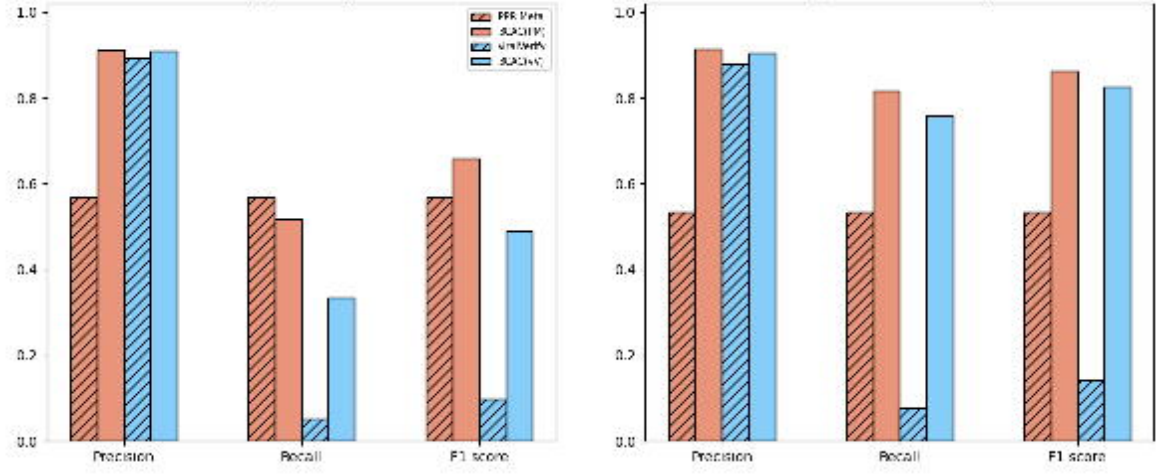

Sim4
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1 score

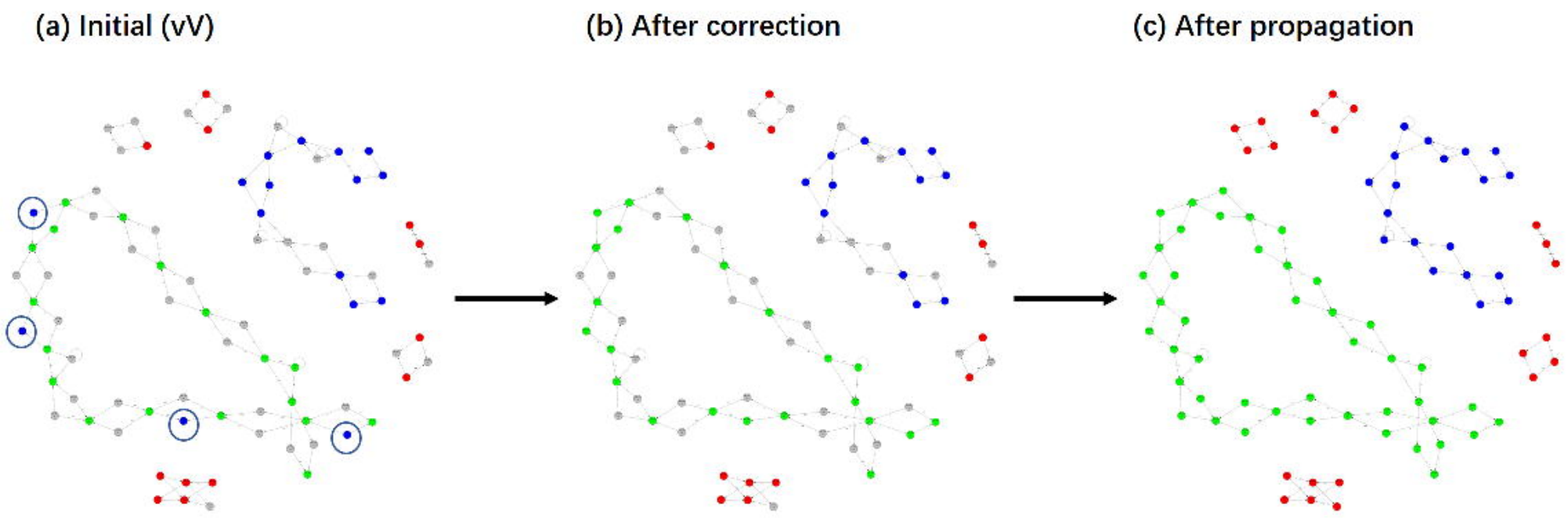

Sim3
Sim4
Recall
Precision
Contig length (kb)
Contig length (kb)
Initial(PM)
3CAC(PM)
Initial(vVI
3CAC(vVI


Sim1
Sim2
Recall
Precision
Contig length (kbl
Contig length (kbl
Initial(PM)
3CAC(PM)
Initial(vV)
3CAC(vV)


Sim1
Sim2
Sim3
Sim4
Recall
1.0
0.8
0.6
0.4
0.2
0.0
Phage
Plasmid
Chromosome


Sim1
Sim2
Sim3
Sim4
Precision
1.0
0.8
0.6
0.4
0.2
0.0
Phage
Plasmid
Chromosome


(a) All contigs
(b) Non-isolated contigs
1.0
0.8
0.6
0.4
0.2
0.0
Precision
Recall
F1 score


(a) Initial (vV)
(b) After correction
(c) After propagation