Comparing and evaluating metagenomic HiFi read assemblers

*Junyeong Ma[1] and Insuk Lee[1]**

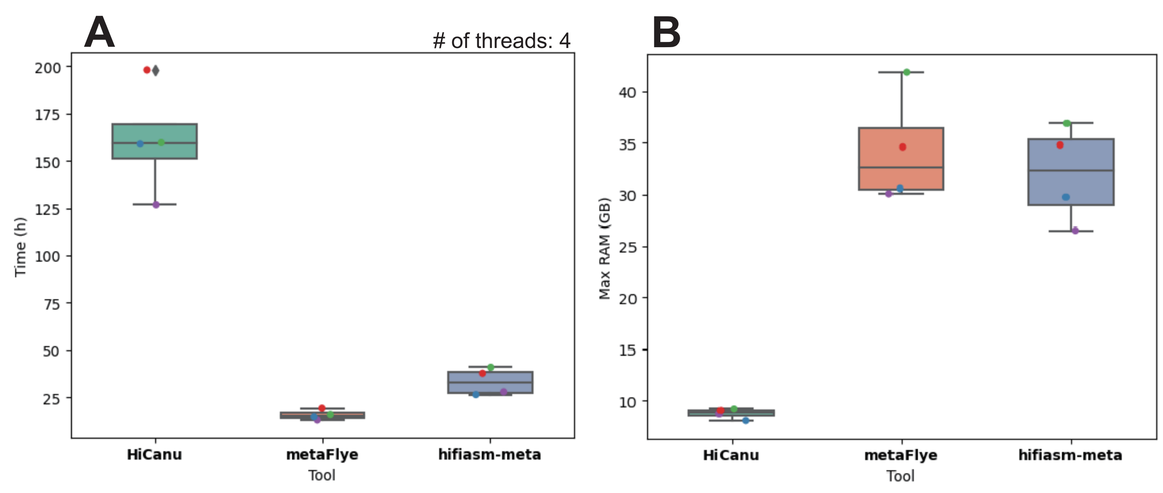
[1] Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Korea

I. Introduction

- Currently, most metagenomic sequencing data are available as short-read sequences. Desptie high accuracy, short read sequencing technologies have an intrinsic drawback in de novo genome assembly due to the limited read length.
- Alternatively, we may use long-read sequencing to overcome the length limitation, but it suffers from poor accuarcy.
- Recently developed PacBio HiFi sequencing[1] was reported to complement the shortcomings of previous long-read sequencing by achieving high accuracy through circular consensus sequencing.
- In this research, we benchmarked three HiFi read assemblers for de novo assembly of metagenomes (HiCanu[2], metaFlye[3], hifiasm-meta[4]) using four public HiFi sequenced samples.

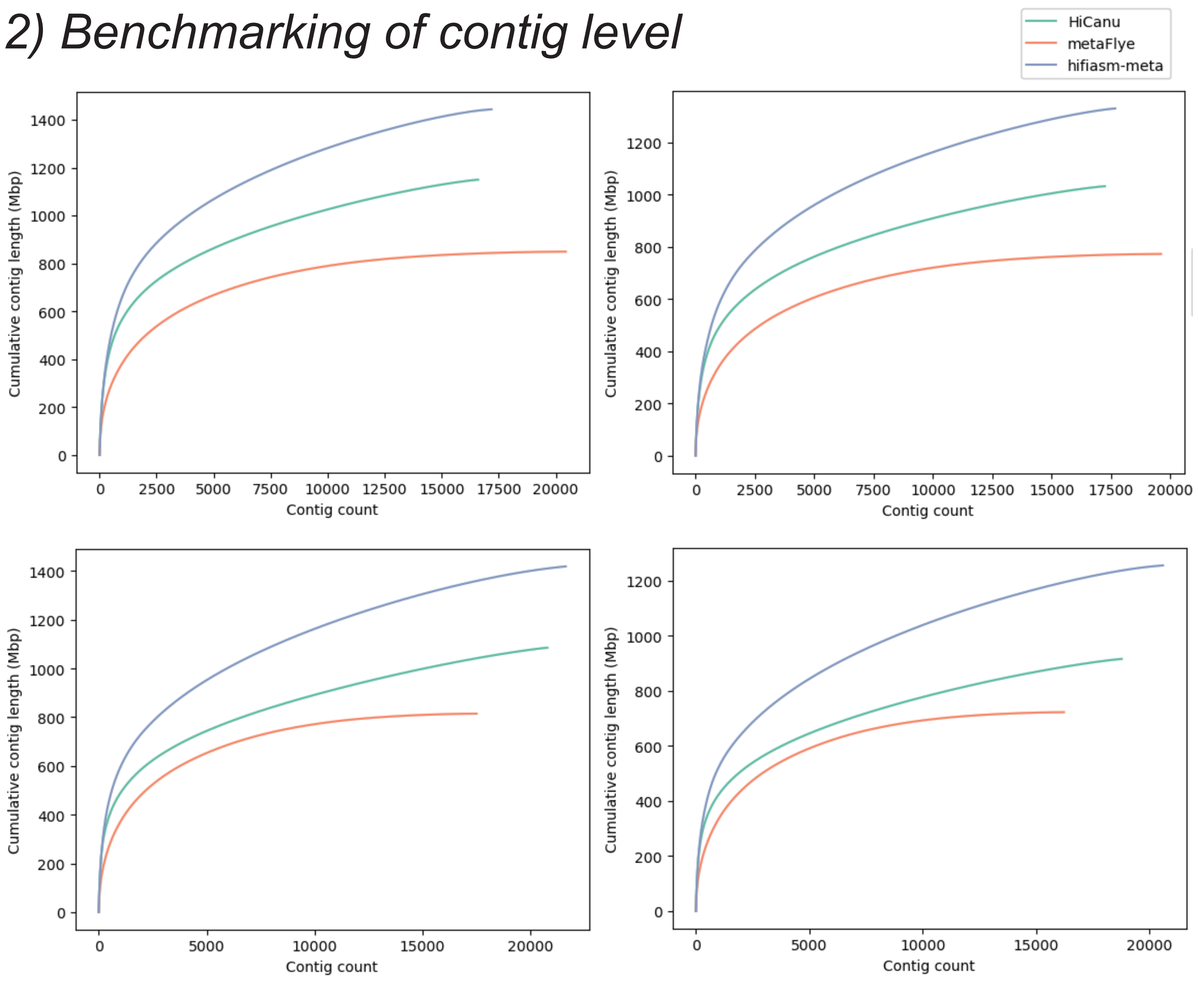
II. Results

1) Benchmarking of computation power

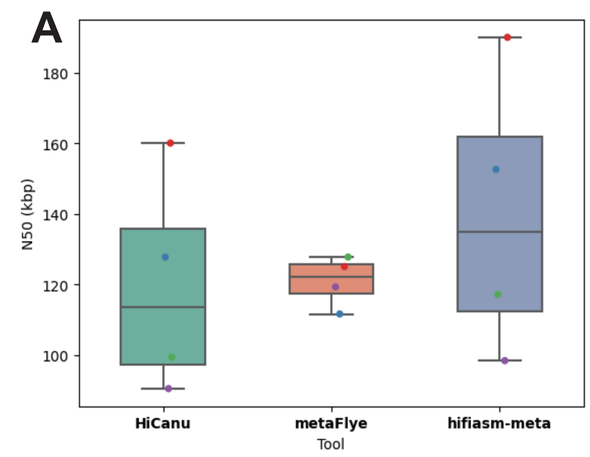


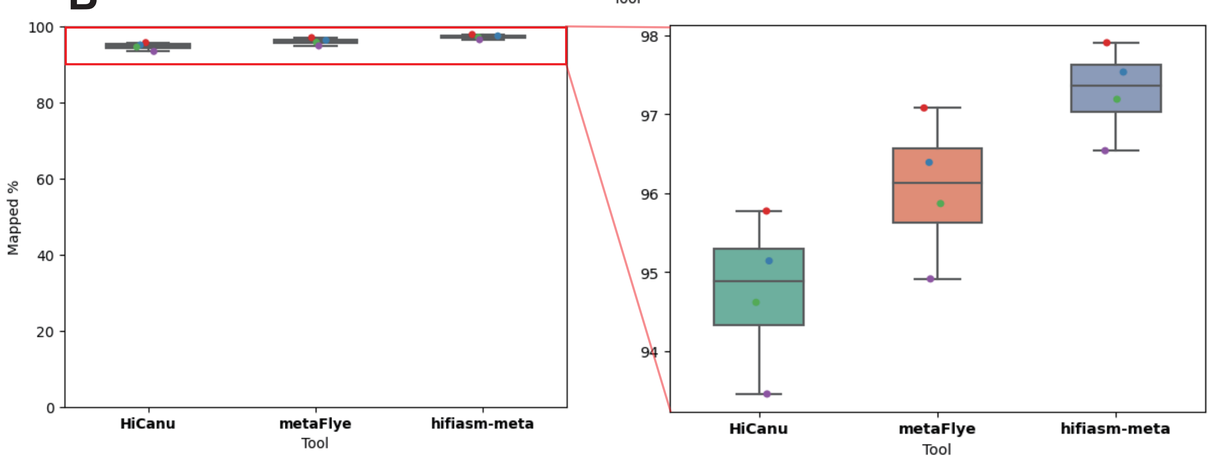
- In terms of time required, metaFlye was the fastest, followed by hifiasm-meta. Compared to the other two tools, HiCanu took a lot of time (Fig. A).
- In terms of the amount of RAM used, HiCanu had the lowest RAM usage. The RAM usage of metaFlye and hifiasm-meta was similar (Fig. B).

2) Benchmarking of contig level



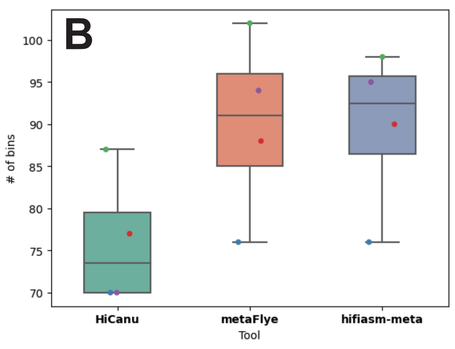
- Through the cumulative contig length graph for four public samples, it was confirmed that hifiasm-meta had the best performance.
- In general, a large number of contigs could be obtained from hifiasm-meta. Not only that, the length of the contig obtained from hifiasm-meta was also long.





- The N50 distribution of hifiasm-meta was relatively larger than that of the other two tools (Fig. A).
- When metagenomic reads were mapped to contigs, there was no significant difference in the mapping ratio. When looking more closely, hifiasm-meta had the highest mapping ratio, followd by metaFlye and HiCanu (Fig. B).

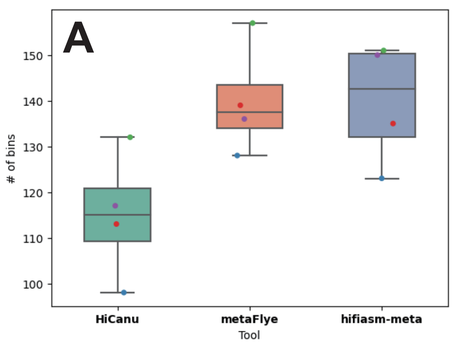
3) Benchmarking of bin level



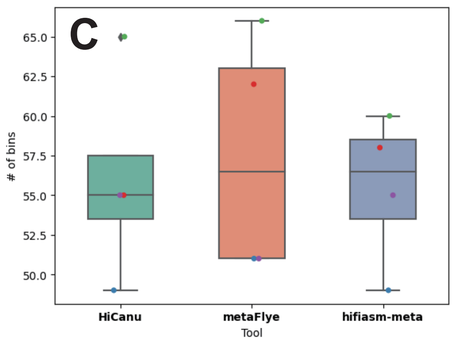
A	Bins
Completeness (%)	>= 50
Contamination (%)	<= 10
Quality score (%)	NA
pass.GUNC	NA

Quality score
: Completeness - 5 * Contamination

C	NC Bins
Completeness (%)	>= 90
Contamination (%)	<= 5
Quality score (%)	>= 50
pass.GUNC	True



B	Bins
Completeness (%)	>= 70
Contamination (%)	<= 5
Quality score (%)	>= 50
pass.GUNC	True



- After binning using metaBAT2, maxBin2, and CONCOCT, bin refinement was performed with metaWRAP. And bin quality control was performed through CheckM and GUNC.
- HiCanu had the smallest number of good quality bins. In the case of metaFlye and hifiasm-meta, the number of bins with good quality was similar, or the number of bins with good quality in metaFlye was higher (Fig. A~C).

III. Conclusions

- HiCanu had the worst performance in terms of time required and the number of bins with good quality.
- When metaFlye and hifiasm-meta were compared, hifiasm-meta performed better at the contig level.
- However, at the bin level, the performance of metaFlye and hifiasm-meta was simliar, or the performance of metaFlye was better.

***Reference**

1) Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology (2019).
2) S. Nurk. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Research. (2020).
3) M. Kolmogorov. et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nature methods. (2020).
4) H. Li. 2019. hifiasm-meta. https://github.com/lh3/hifiasm-meta. (2021).