

Assembly of long, error-prone reads using repeat graphs



Abstract

With the emergence and the spreading adoption of long-read technologies, the assembly of genomes has improved, which may also become the “gold-standard” for de novo assemblies. Current long read sequencing techniques still have high error rates which makes it more difficult to align these reads. Up to date long read assemblers are still not able to resolve all repeating regions correctly. Especially segmental duplications, long and highly homologous sequences resulted from duplications, are still problematic to resolve correctly. While genomic repeats can be better resolved using long reads, assembly with them is still challenging and not straightforward due to their error-prone nature. Here we present Flye a de novo assembler for long-error prone reads, by creating a precise repeat graph, built in a new manner using so called disjointigs. Flye could achieve two times better contiguity for the assembly of a human Oxford Nanopore test dataset in combination with short read Illumina data in contrast to the state of the art assembler Canu. In the created repeat graph many segmental duplications are represented from which the simple ones are already resolved by the algorithm. Our assembler shows that a genome can be accurately assembled by repeat characterization using repeat graphs. This information can also help in improving existing assemblies. With the presented algorithm, a possibility is provided to improve the de novo assembly of a genome.

Background

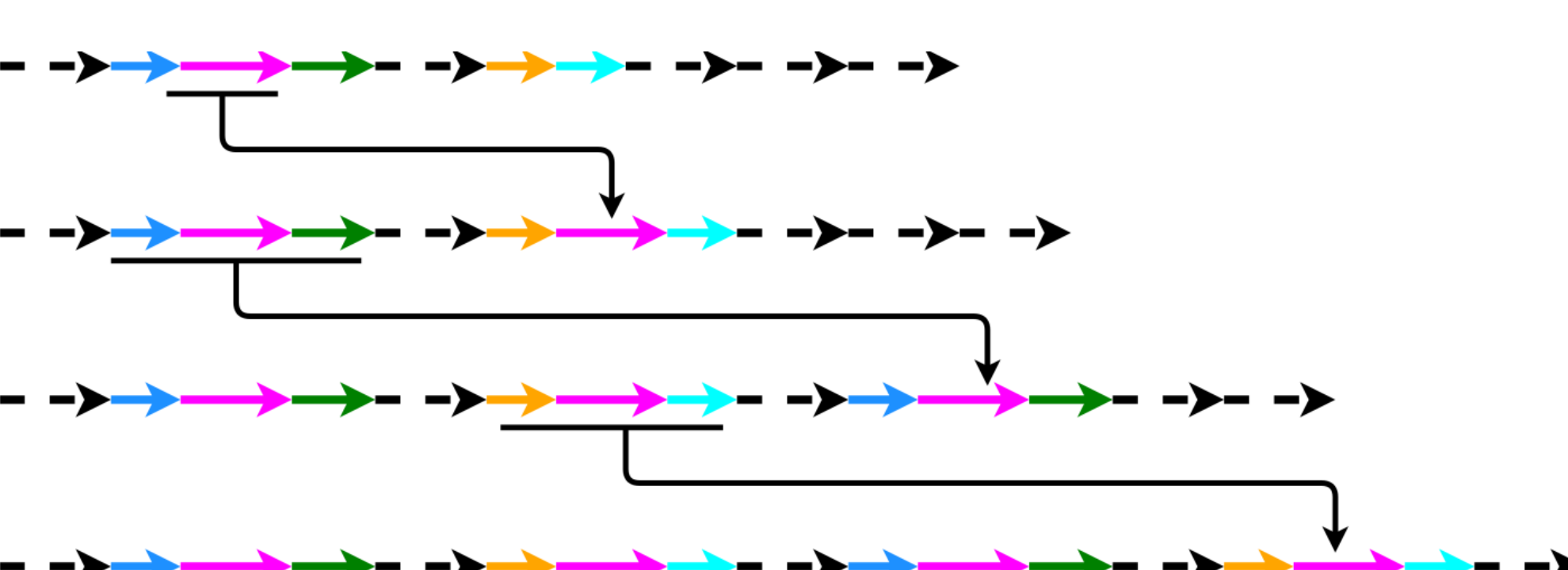
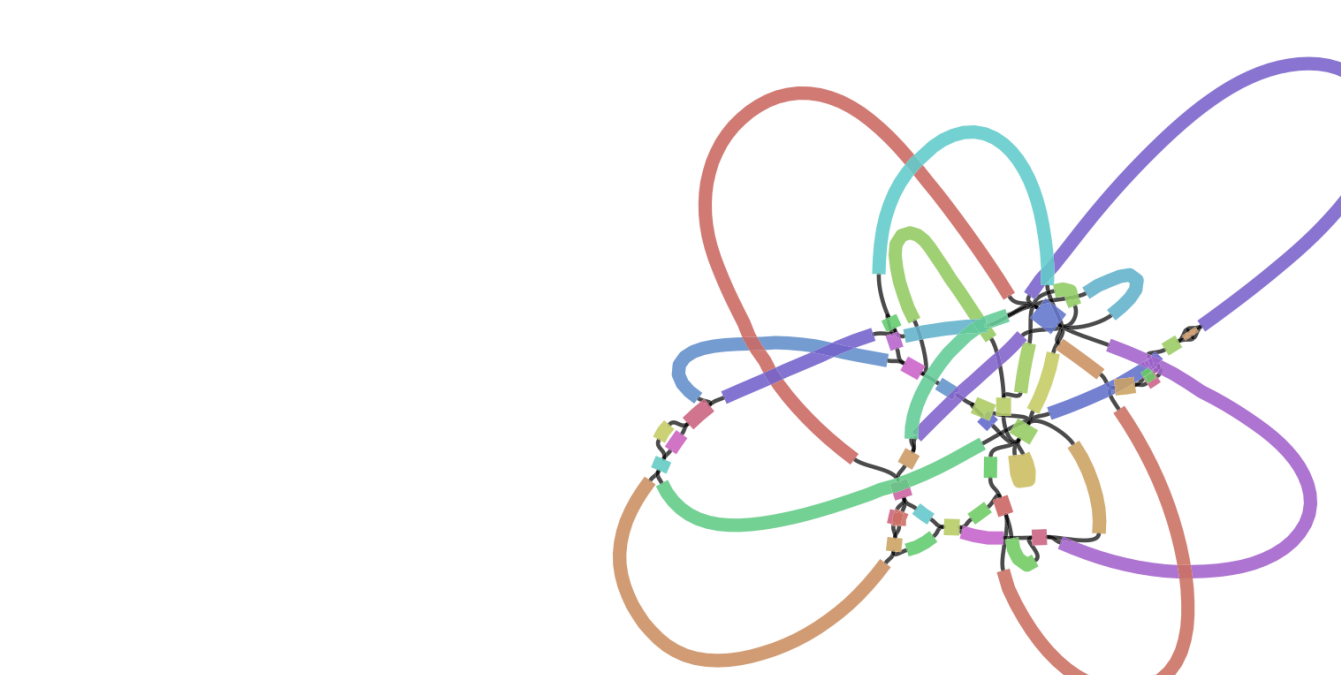
- Different assemblers, different graph structures (De-Bruijn, Overlap-layout, String)
- Segmental duplications (SD)

- Assembly graph is tangled if SD/mosaic repeats are present → fragmentation


Figure 1: How segmental duplications arise (modified from [2])

- Small differences between repeat copies are harder to resolve using error prone-reads

Aim

- generate an algorithm, which is able to:
 - resolve repeating regions
 - assemble the long error-prone reads correctly

Methods

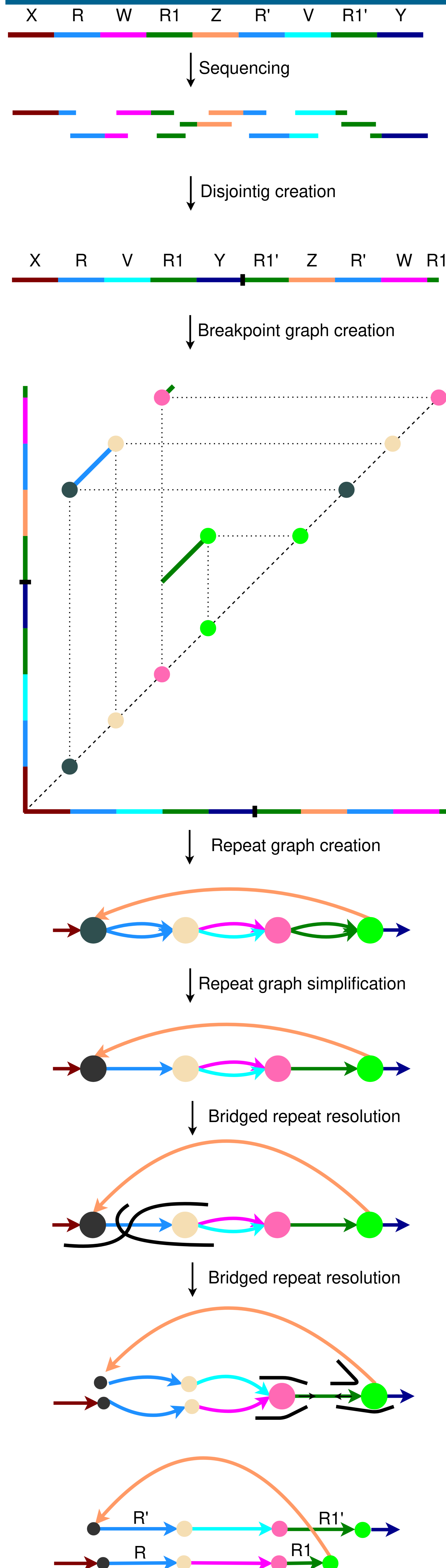


Figure 3: Repeat graph creation and repeat resolution (modified from [1])

Results

- Flye is able to create more contiguous assemblies than other state of the art sequencing algorithms

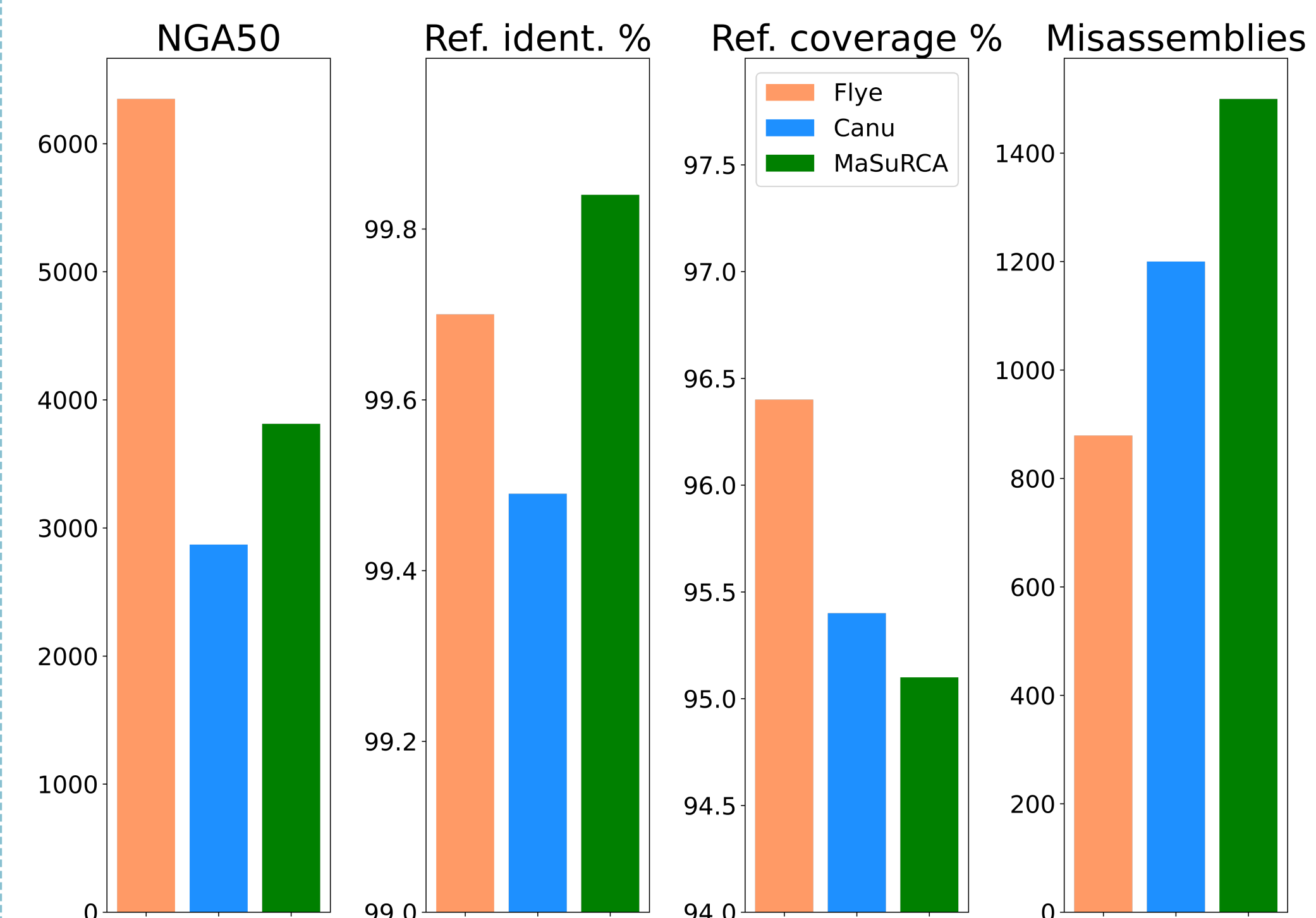


Figure 4: Comparison of Flye assembly against the state of the art assemblers Canu and Masurca. Assembly of a human Oxford Nanopore long read dataset. Comparing the NGA50, reference percentage identity, reference percentage coverage and the number of misassemblies (from left to right). (Own figure from table 1 in [1])

- Refinements of the Flye algorithm lead to improvements in assembly contiguity → better results only by algorithmic improvements

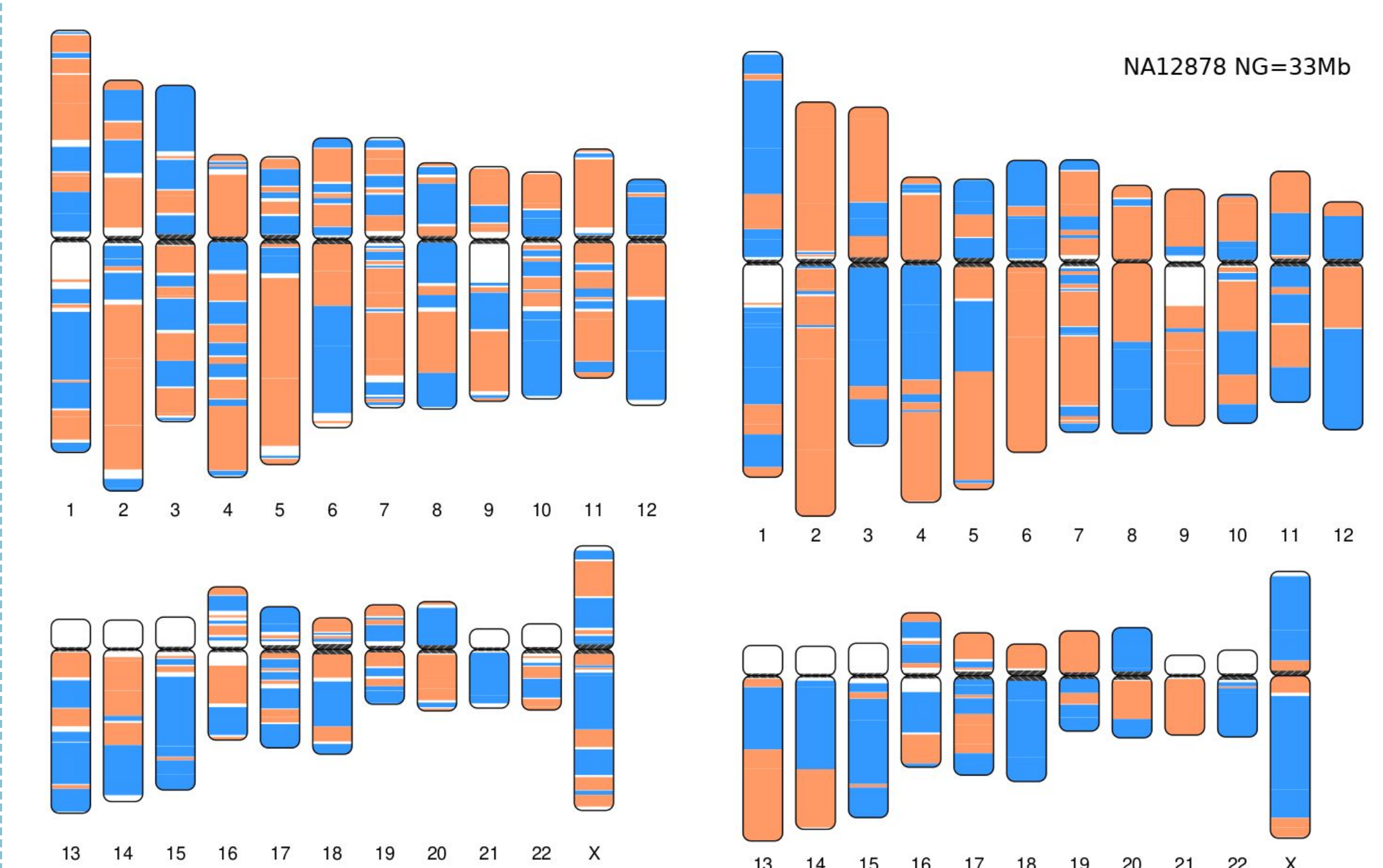


Figure 5: Contiguity comparison for the assembly of the same Oxford Nanopore human dataset using different versions of the Flye assembler. Different colors correspond to different contigs. (M. Kolmogorov, personal communication, June 30, 2021)

References

1. Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37(5), 540-546. <https://doi.org/10.1038/s41587-019-0072-8>
2. Pevzner, P. A., Pevzner, P. A., Tang, H., & Tesler, G. (2004). De novo repeat classification and fragment assembly. *Genome Research*, 14(9), 1786-1796. <https://doi.org/10.1101/gr.2395204>
3. Lin, Y., Yuan, J., Kolmogorov, M., Shen, M. W., Chaisson, M., & Pevzner, P. A. (2016). Assembly of long error-prone reads using de Bruijn graphs. *Proceedings of the National Academy of Sciences*, 113(52), E8396-E8405. <https://doi.org/10.1073/pnas.1604560113>