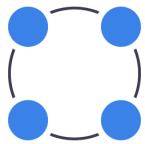
Long read mapping improvements for Flye assembler

Evgeny Polevikov

Scientific advisor: Mikhail Kolmogorov
University of California San Diego



Sequencing technologies

- **Short-read** sequencing technologies (read length is $\approx 25 - 1100$, error rate is $\approx 1 - 2\%$):
  - Illumina
  - Roche 454
  - Ion Torrent
- **Long-read** sequencing technologies (read length is $\approx 1000 - 200\,000$, error rate is $\approx 10 - 30\%$):
  - Pacific Biosciences
  - Oxford Nanopore

Genomy assembly

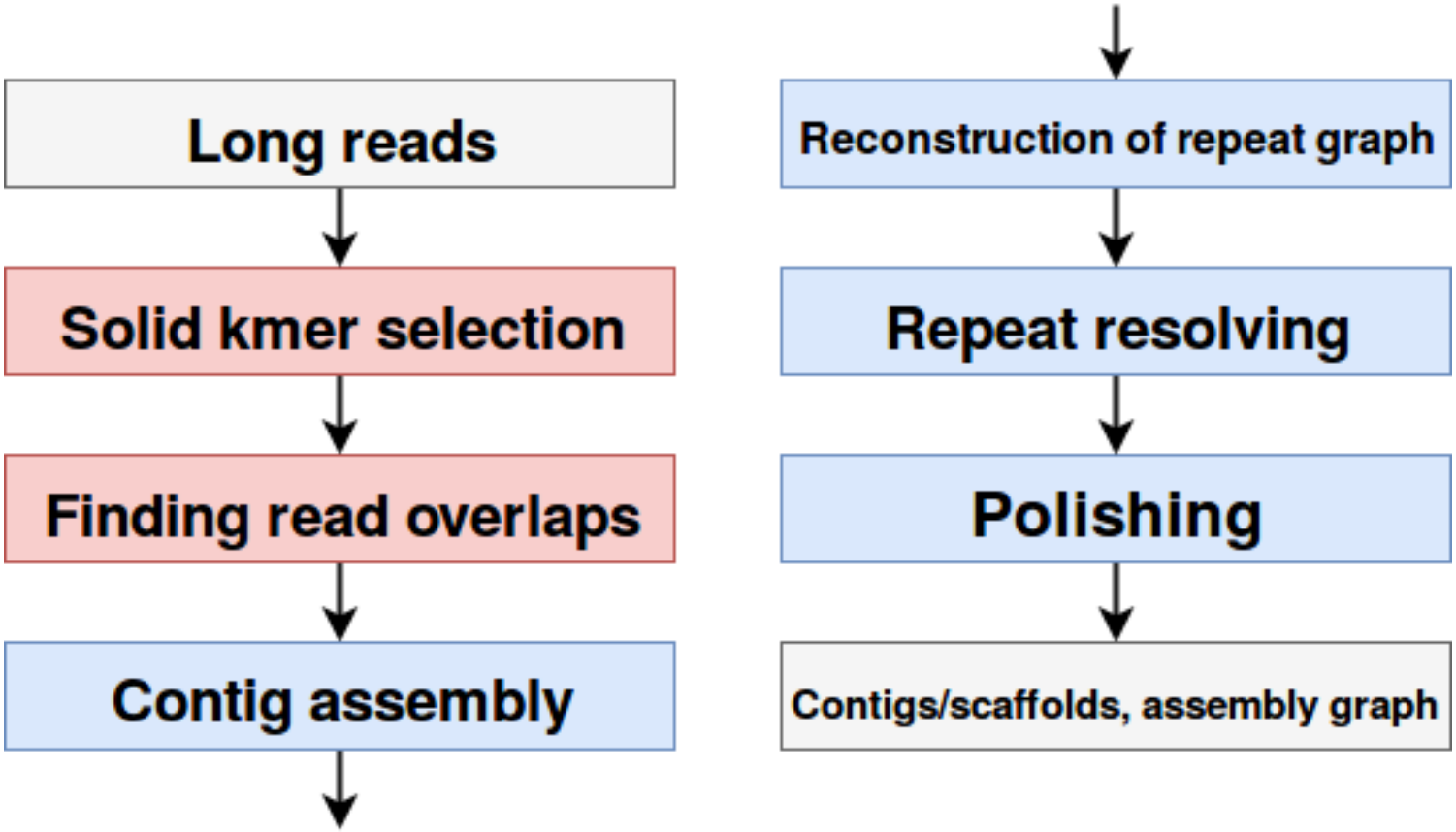
- The primary challenge to all assembly algorithms is repeats. Long reads allow to resolve them more efficiently comparing with shorts reads.
- Long-reads assemblers:
  - Canu
  - Falcon
  - Flye
  - HINGE
  - Miniasm
  - ...

The goal of the project

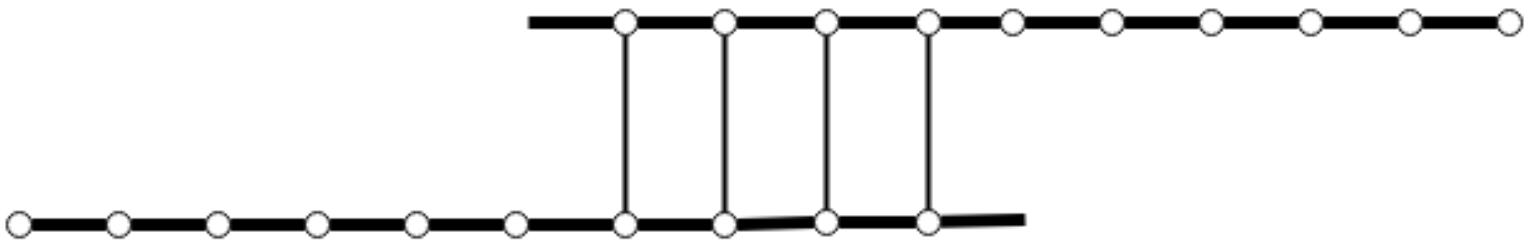
- **Flye** (Kolmogorov M, et al. bioRxiv 2018)
- **minimap2** (Li H, arXiv 2017)

The goal of the project is to incorporate **minimap2** into **Flye** in order to reduce memory usage bottleneck and improve assembly accuracy.

The Flye assembly pipeline



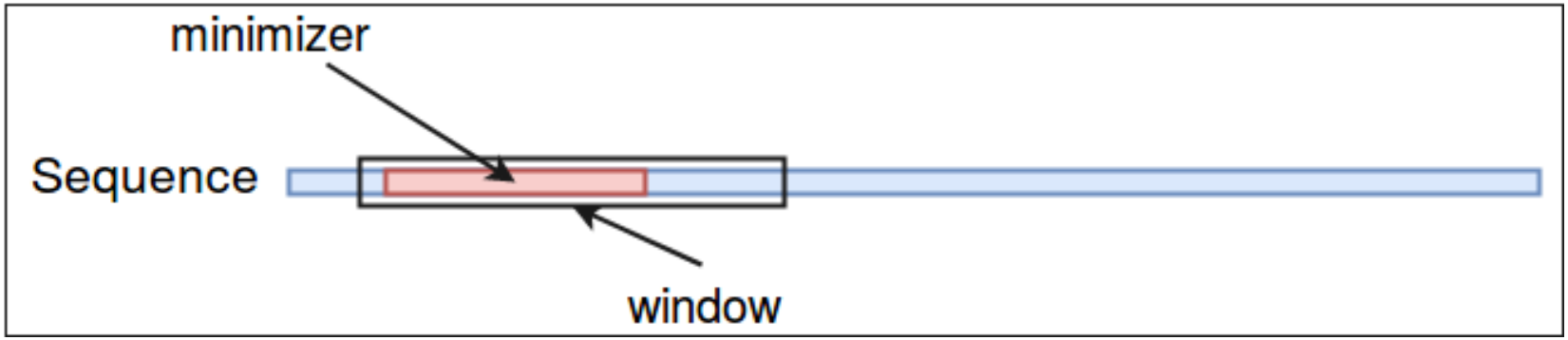
Finding read overlaps using solid kmers



Shortcomings of the current approach:

- Positions of overlaps are less accurate comparing with those we can find using alignment
- An index for the solid kmers takes a lot of memory
- Low flexibility of the parameters choice

Minimap2 approach for finding overlaps



- Minimizer of a sequence is a minimal k-mer in a window of size w.
- In order to find overlaps minimap2 constructs a set of minimizers for target and query sequences and finds hits between them.

The results obtained

E.coli dataset: input reads – 203Mb, genome size – 5Mb, coverage – 40x.

	Memory (assembly module), Gb	Memory (total), Gb	# contigs
minimap2	1.9	–	–
flye-vertex	2.1	5.6	1
flye-minimap	2.2	5.6	1

The results obtained

S. cerevisiae dataset: input reads – 367Mb, genome size – 12Mb, coverage – 31x.

	Memory (assembly module), Gb	Memory (total), Gb	# contigs	N50	# misassembled contigs
minimap2	3.1	–	–	–	–
flye-vertex	4.9	5.0	33	605k	11
flye-minimap	3.4	4.7	26	748k	8

The results obtained

X. oryzae dataset: input reads – 1.1Gb, genome size – 4.8Mb, coverage – 220x.

	Memory (assembly module), Gb	Memory (total), Gb	# contigs	N50	# misassembled contigs
minimap2	8.0	–	–	–	–
flye-vertex	5.2	10.7	5	4.8M	0
flye-minimap	8.0	11.8	4	2.8M	2

Summary

- The algorithm of finding overlaps in Flye was substituted by seed-chain-align procedure using minimap2 API.
- But it did not decrease memory usage significantly.
- Nonetheless, another ways of improvements are possible: e.g saving of an index to a disk and finding overlaps using the constructed index by parts.