GoldRush: A *de novo* assembler for long reads with linear time complexity

Johnathan Wong, Vladimir Nikolic, Lauren Coombe, Emily Zhang, René L. Warren, and Inanc Birol

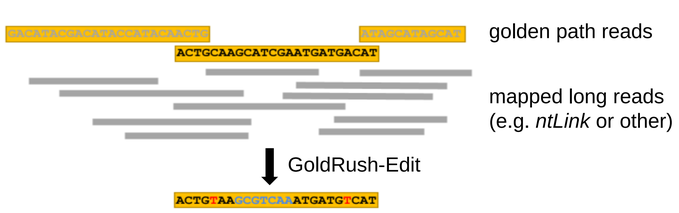
Introduction

State-of-the-art *de novo* long read assemblers

- Overlap-Layout-Consensus paradigm
- High RAM cost

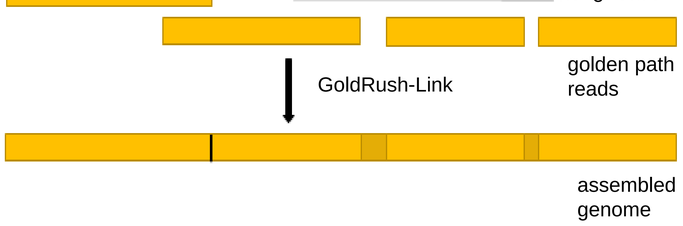
GoldRush

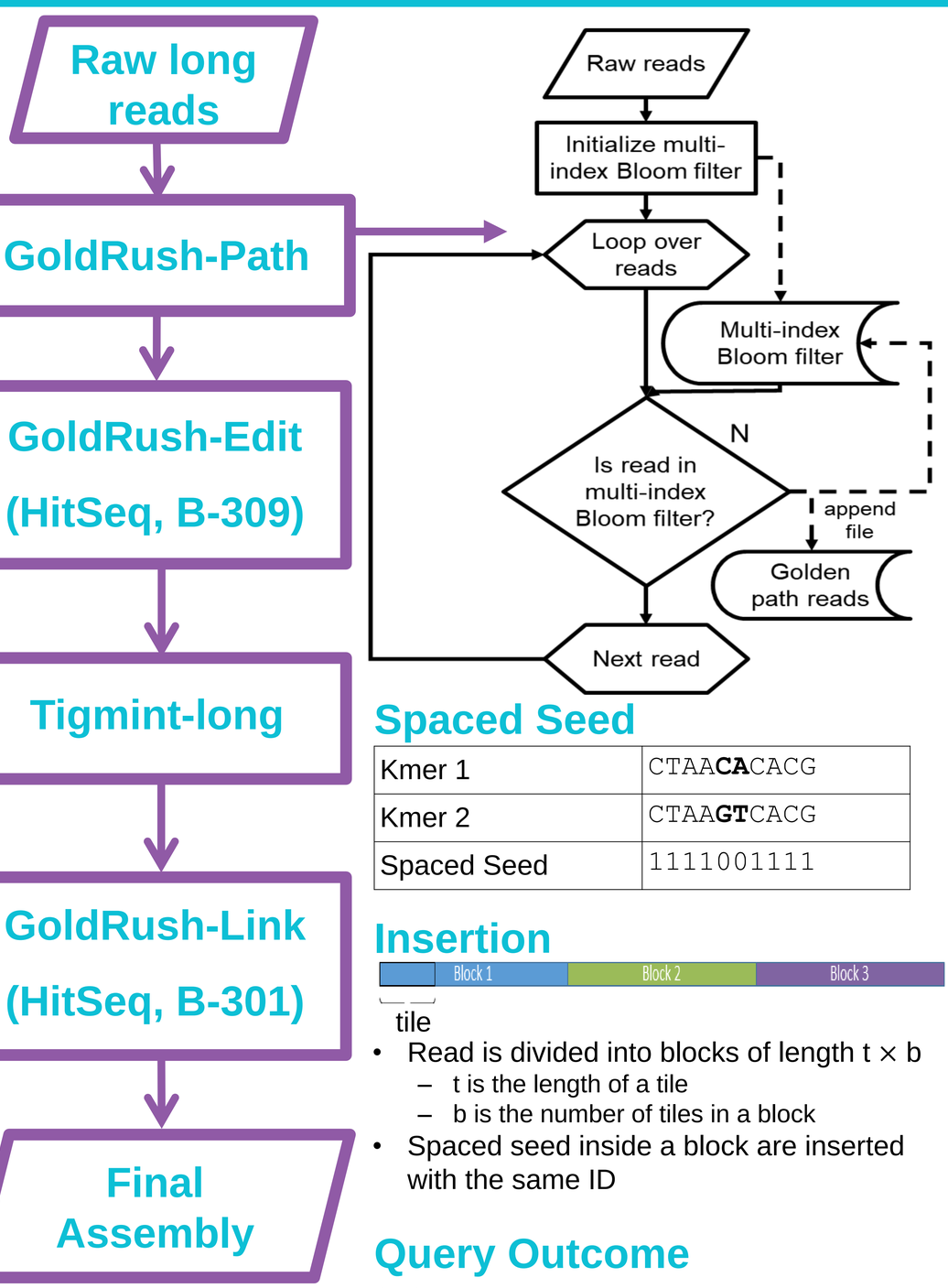
- Linear time algorithm
- Low RAM cost



Detects and removes misassemblies

- Identify and cut chimeric golden path reads





Spaced Seed

Kmer 1	CTAA**CA**CACG
Kmer 2	CTAA**GT**CACG
Spaced Seed	1111001111

Insertion

- Read is divided into blocks of length t × b
  - t is the length of a tile
  - b is the number of tiles in a block
- Spaced seed inside a block are inserted with the same ID

Query Outcome

Insert:

Read									
ID	73	20	15	1	1	1	223	5000	195
Assignment	false	false	false	false	false	false	false	false	false

Trim and insert:

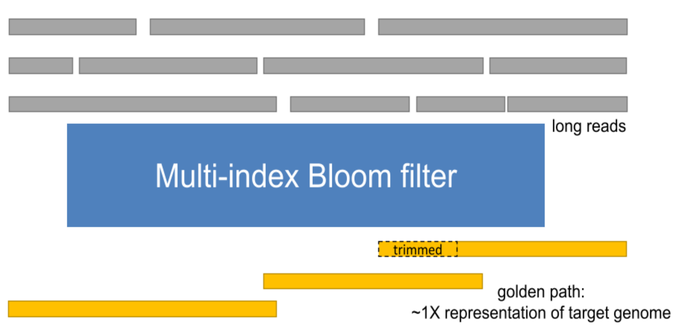
Read									
ID	1	1	1	1	2	2	2	20	23
Assignment	true	true	true	true	true	true	true	false	false

Read		
ID	2	20	23
Assignment	true	false	false

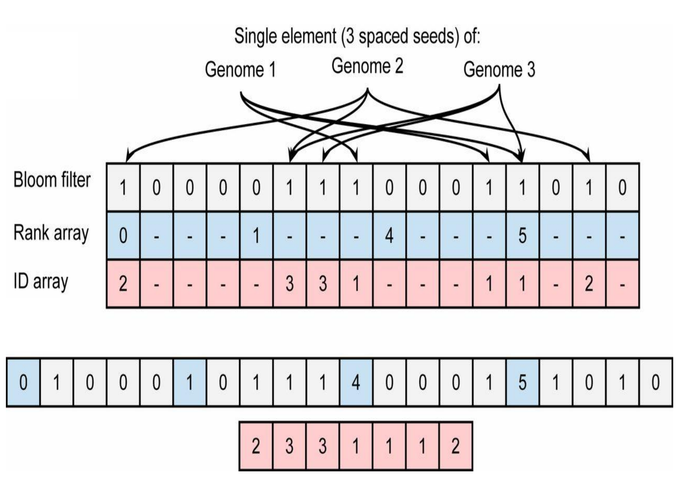
Skip

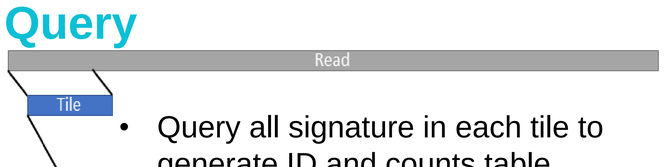
Read						
ID	1	1	1	1	2	2
Assignment	true	true	true	true	true	true

Golden Path



Multi-index Bloom filter





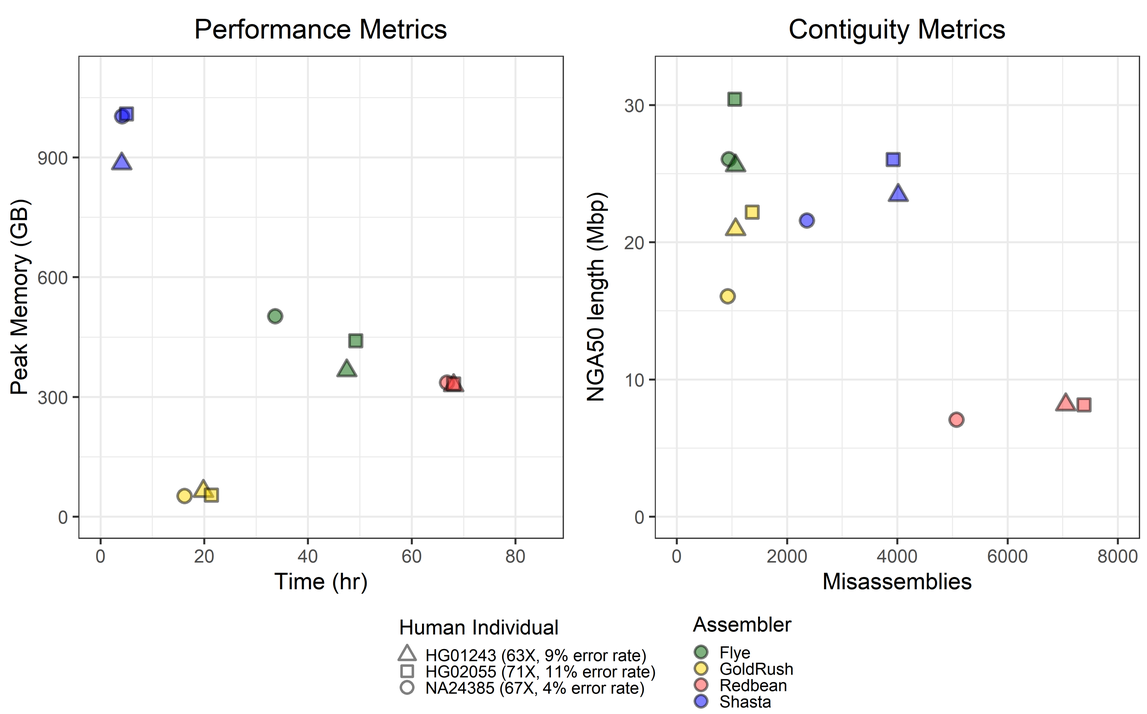
Query

Best hits

Read									
ID	1	1	1	5	2	2	2	20	23
Counts	900	756	300	500	8	100	200	9	5
Assignment	true	true	true	true	false	true	true	false	false

- Preliminary assignment is based on a threshold x (10)
- If a tile is assigned, then the tile is found in the golden path

Results



Conclusions

- Novel *de novo* long read assembly algorithm
- Lowest RAM cost
- High contiguity and low misassemblies

Software Availability

https://github.com/bcgsc/goldrush

References

Coombe L, et al. 2021. LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics* 22:2021.06.17.448848. DOI: 10.1186/s12859-021-04451-7.
Chu J, et al. 2020. Mismatch-tolerant, alignment-free sequence classification using multiple spaced seeds and multiindex Bloom filters. *Proceedings of the National Academy of Sciences* 117:2020.07.08. DOI: 10.1073/pnas.1903436117.

Contact

jowong@bcgsc.ca

Funding

Genome BritishColumbia
GenomeCanada
National Institutes of Health