Mapping noisy long-reads with multi-index Bloom Filter: miBF-mapper

T. Murathan Goktas[1,2], Vladimir Nikolic[1], Jonathan Wong[1], Ka Ming Nip[1] Lauren Coombe[1], Rene Warren[1], Inanc Birol[1,2]

[1] Canada's Michael Smith Genome Sciences Centre, Vancouver, Canada
[2] Bioinformatics Program, University of British Columbia, Vancouver, Canada

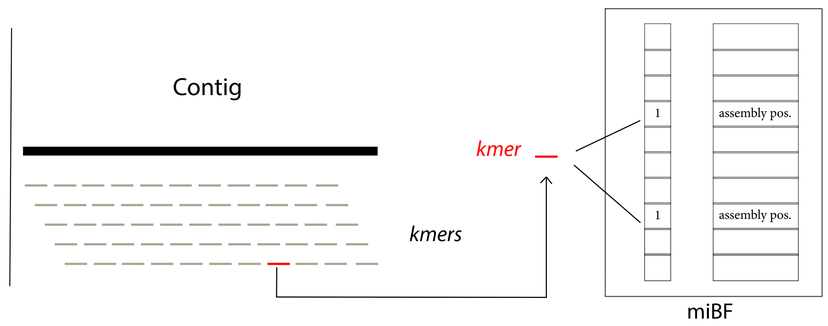
INTRODUCTION

- multi-index Bloom filter (miBF) is an extension of Bloom filter (BF) with an additional ID array for each populated element in BF.
- miBF outpeformed competitors in short read classification in 2020.[1]
- Here we investigated and benchmarked its utility in long-read mapping.

METHODS

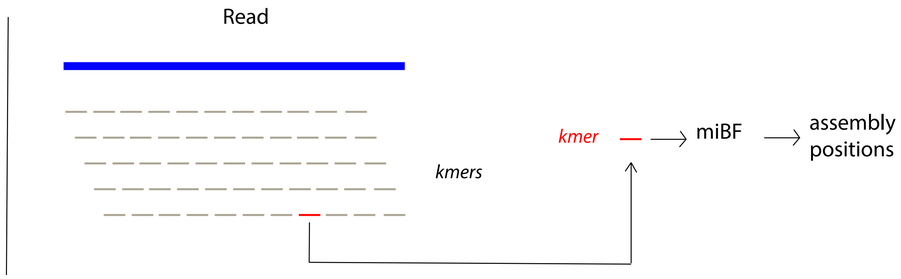
1. Build multi-indexed Bloom Filter with assembly

- miBF is built from assembly/reference genome.
- ID's inserted for each kmer is the assembly position of the kmer.



2. Query reads on miBF

- All kmers of read are queried on miBF and the results are stored.
- Stored results are then sorted by relative positions and strand.



RESULTS

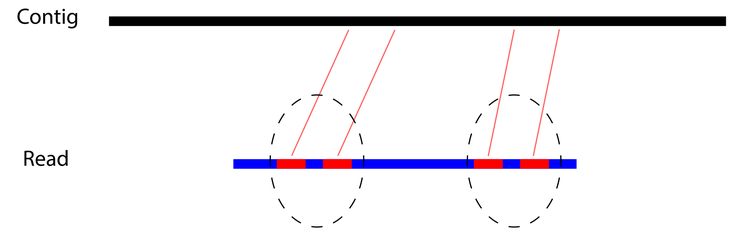
	C. elegans		*H. sapiens**	
	miBF-mapper	**minimap2[2]**	**miBF-mapper**	**minimap2**
Accuracy**	99.9%	99.9%	92.6%	93.4%
Runtime	2m:02s	10s	36m:41s	2m:5s
Max Memory (GB)	2	1	75	10

Reads: 50k Nanosim[3] simulated long-reads with 11% error rate
*GRCh38
**Correct mappings have <10% overlap with the correct region

miBF-mapper parameters:
-kmer length: 21bp
-#hash functions: 5

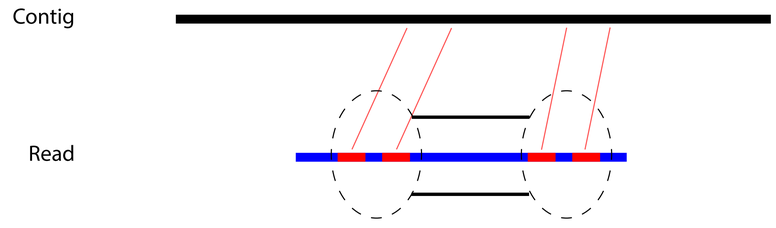
3. Cluster hits

- Kmer hits are clustered if they have same relative positions and close to each other on the read sequence.



4. Chain clusters

- Close clusters are merged to generate the most confident and long chain.
- That chain is reported to output as the mapping.



CONCLUSIONS

- miBF-mapper is comparable in accuracy to best tools in domain and requires more computational resource.
- miBF data structure is usable in various genomics applications.
- As a maximum of #hash functions positions can be stored in miBF, few positions of repetitive kmers can be fetched from miBF in a query. This causes performance degradation in mapping repetitive regions with miBF.
- miBF is more suitable for applications where rare kmers give more signal, like classification applications.

REFERENCES

1. Chu J, et. al. (2020) PNAS; 117 (29) 16961-16968.
2. Yang C, et. al. (2017) GigaScience; Volume 6, Issue 4
3. Li H, (2018) Bioinformatics; Volume 34, Issue 18, 3094–3100

PRESENTER

T Murathan Goktas

@murathangoktas_
tgoktas@bcgsc.ca

https://github.com/MurathanGoktas/miBF-mapper

ACKNOWLEDGEMENT

GenomeCanada
Genome BritishColumbia
BC CANCER