

MindTheGap: integrated detection and assembly of short and long insertions



Guillaume Rizk¹, Anaïs Gouin¹, Rayan Chikhi² and Claire Lemaître¹

¹ Inria/IRISA GenScale, Campus de Beaulieu, 35042 Rennes cedex, France.

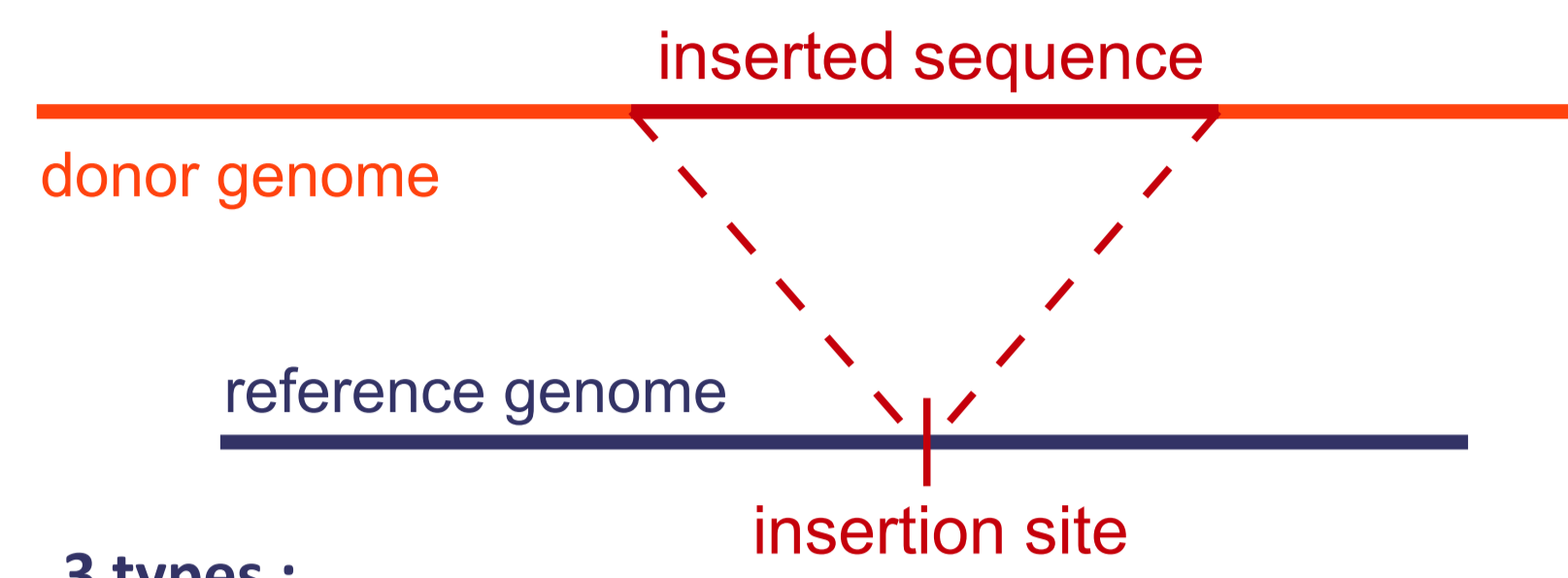
² Department of Computer Science and Engineering, Pennsylvania State University, USA.
guillaume.rizk@inria.fr, claire.lemaître@inria.fr

PENN STATE



Motivations

Insertion Variants – definition

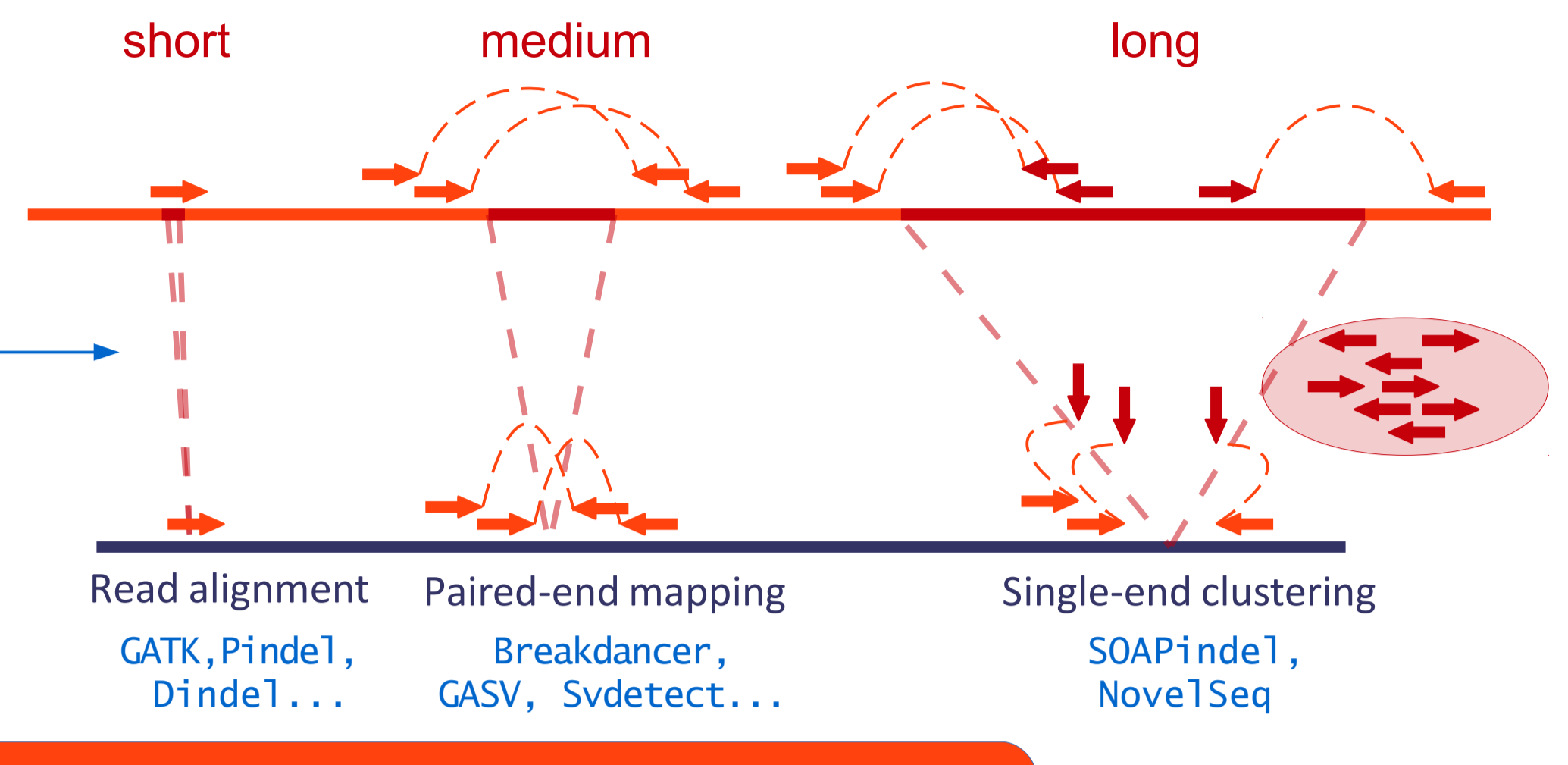


3 types :

- *novel* insertion
- *duplicative* insertion
- *transposition*

Insertion Variants – detection in short read data

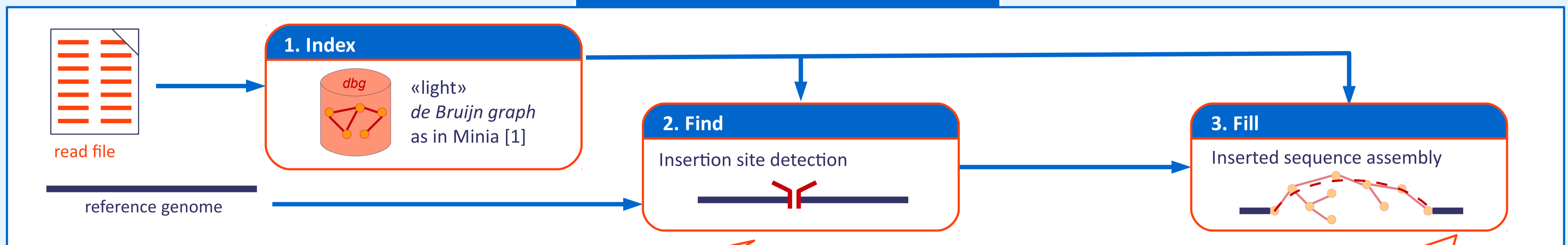
1. Mapping the reads to the reference genome
2. Insertion site detection
Several methods depending on the insertion size
3. Assembly of the inserted sequence
Few methods for medium/large insertions
Local assembly of mate reads → limited size
Assembly of unmapped reads → only novel insertions



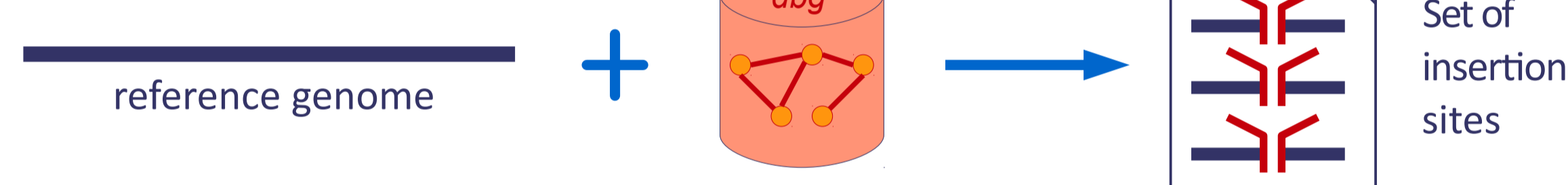
No tool for all types and all sizes insertions
Especially strong need for tools for long insertions (> insert-size)

Methods

MindTheGap



Find module



Scanning the reference genome and testing existence of kmers in the donor *dbg*

Homozygous insertions generate *gaps* of size < k

k-1 missing k-mers

11111100000111111111111111

GTATTACTATGCTATCTATTATTTA S_r

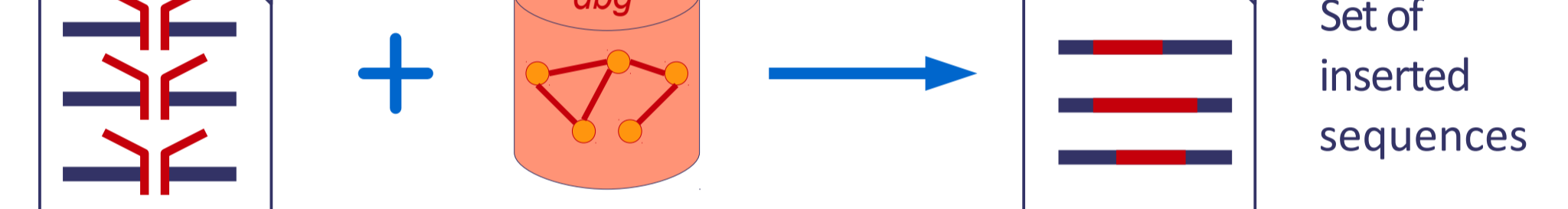
GTATTACTATG...insertion...CTATCTATTATTTA S_i

Heterozygous insertions generate overlapping *branching* kmers

GTATTACTATGCTATCTATTATTTA

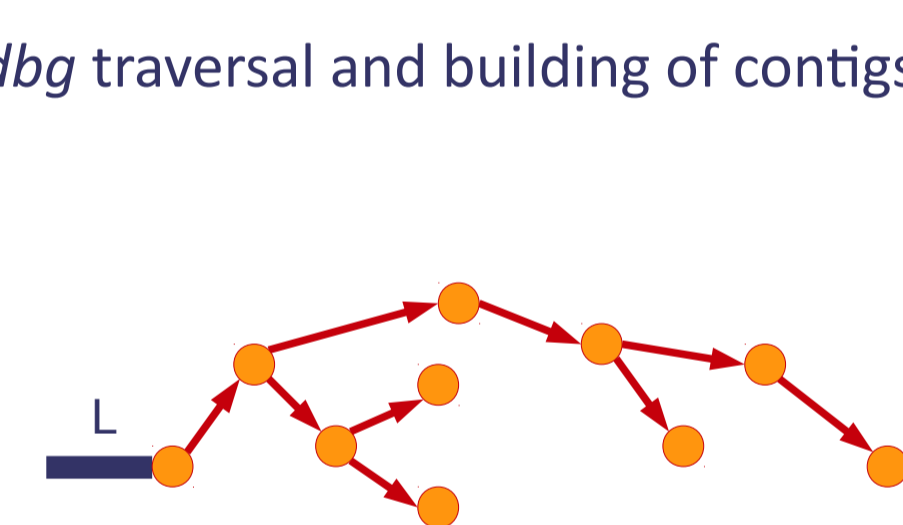
GTATTACTATGCTATCTATTATTTA

Fill module

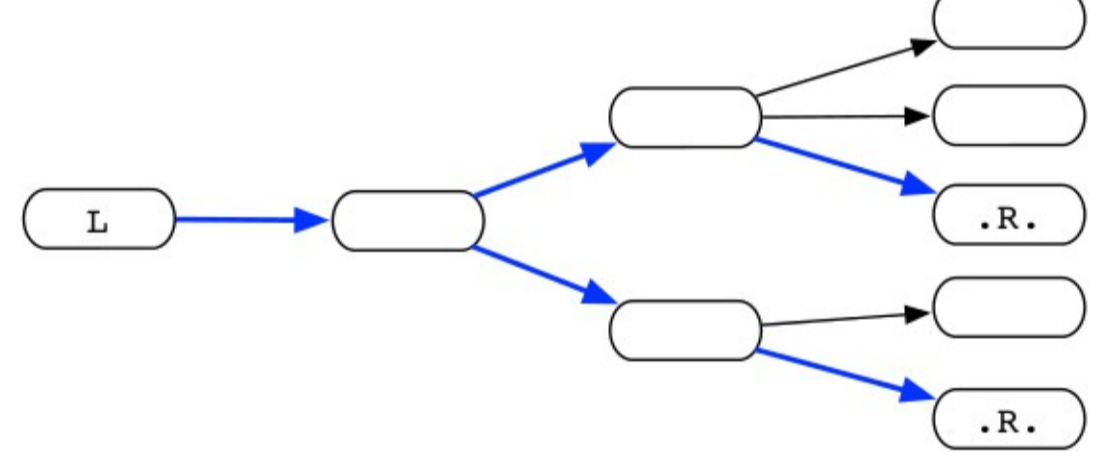


De novo assembly starting from *Left* kmer, searching for *Right* kmer in the contig graph

dbg traversal and building of contigs



In the graph of contigs, finding all paths from L to R

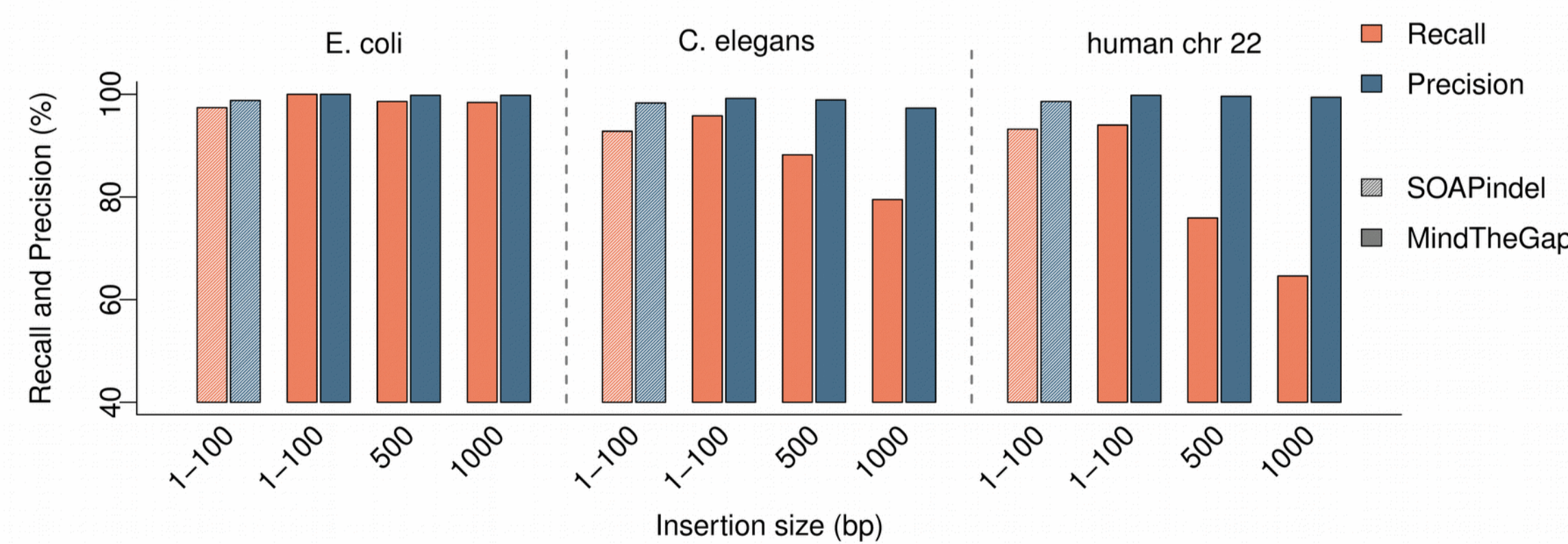


Results

Simulated data:

- Insertions : random deletions in real genomes (ref)
- Donor reads:
 - wgsim (2x100 bp, 40x) on initial genome
 - real *C. elegans* SRX026594 (70x)

Comparison with SOAPindel [2] (limited to 1-100 bp insertions).



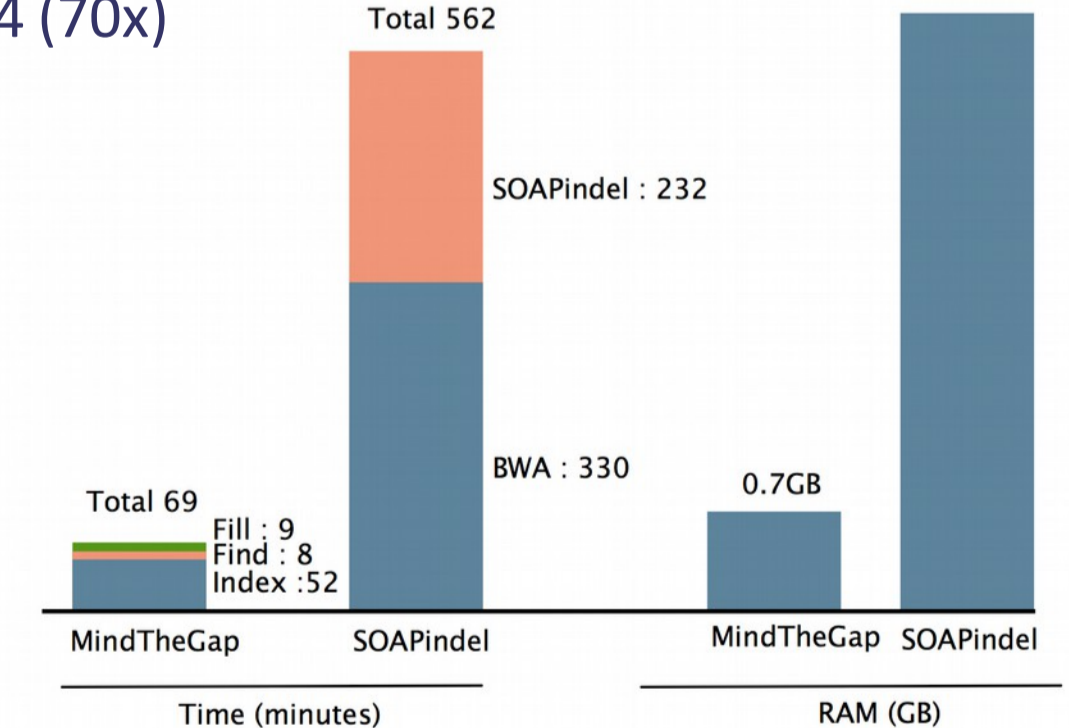
1 Kb insertions	Recall (%)		Precision (%)		N sim.	Find module		Fill module	
	TP	FP	TP	FP		TP	FP	TP	FP
<i>E. Coli</i> simulated dataset	98.4	99.8	500	499	0	492	1		
<i>C. elegans</i> simulated dataset	79.5	97.3	1000	992	0	795	22		
<i>C. elegans</i> real reads, simulated insertions	81.1	-	1000	980	-	811	-		
Human chromosome 22 simulated dataset	64.6	99.4	1000	1000	0	646	4		
Heterozygous insertions									
<i>C. elegans</i> simulated dataset	59.9	93.4	1000	807	11	599	42		
Human chromosome 22 simulated dataset	35.5	89.0	1000	816	28	355	44		

Real «big» data :

- Human NA12878, 2.8 G reads ~100x
- Real insertions detected in fosmid data [3]
- 11/23 detected insertion sites
- 2 insertions of 4,137 and 6,729 pb correctly assembled

Faster and with few memory

On the real *C. elegans* dataset SRX026594 (70x)



On the human «big» dataset (280 Gbp): <14 GB memory and 45 hours

MindTheGap:

- ★ The only tool to assemble large insertions (>1Kb)
- ★ Integrated and unified method for all types and all sizes insertions
- ★ Independent of mapper and insert-size
- ★ Time/memory scalable on big datasets
- ★ Can be used as a gap-filler in assembly pipelines

References:

- [1] R. Chikhi & G. Rizk. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms Mol Biol*, 2013, 8: 22.
- [2] S Li, R. Li, H. Li, J. Lu, Y. Li, L. Bolund, M. H. Schierup & J. Wang. SOAPindel: efficient identification of indels from short paired reads. *Genome Res*, 2013, 23: 195-200.
- [3] S. Kim, P. Medvedev, T. A. Paton & V. Bafna. Reprever: resolving low-copy duplicated sequences using template driven assembly. *Nucleic Acids Res*, 2013, 41: e128.

Download and use MindTheGap

<http://mindthegap.genouest.org>

C++, Open Source, released under A-GPL license

Publication: G. Rizk, A. Gouin, R. Chikhi & C. Lemaître. **MindTheGap : integrated detection and assembly of short and long insertions.** *Bioinformatics*, 2014.

Get back to this poster :

<http://tiny.cc/mtgposter>

