1st International Conference on Algorithms for Computational Biology
AlCoB 2014
Tarragona, Spain, July 1-3, 2014

Mapping-free and Assembly-free Discovery of Inversion Breakpoints from Raw NGS Reads

Claire Lemaitre[1], Liviu Ciortuz[2] and Pierre Peterlongo[1]

[1] Genscale, IRISA/Inria in Rennes, France

[2] Faculty of Computer Science of Iasi, Romania

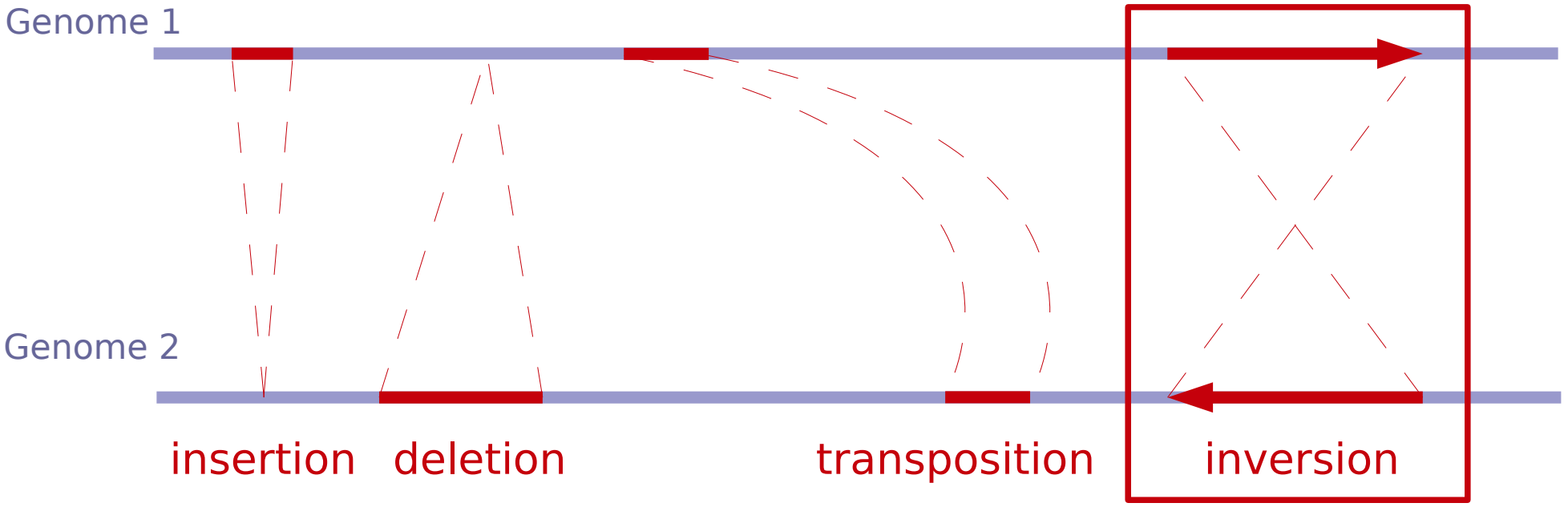
UMR IRISA

GENSCALE

informatiques mathématiques Inria

Structural Variants

- Large mutations of the genome



- Impacts on evolution and disease

  more bp involved in SV than in SNPs

- Next Generation Sequencing (NGS):

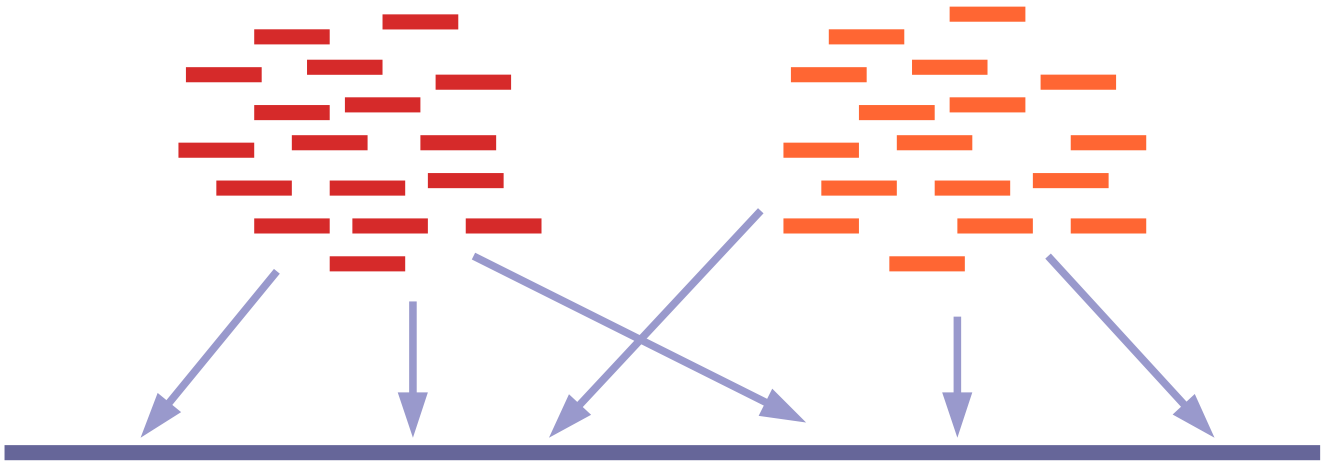
  sequencing whole genomes of several individuals

Structural Variant detection

- NGS: millions of small reads

Structural Variant detection

- NGS: millions of small reads



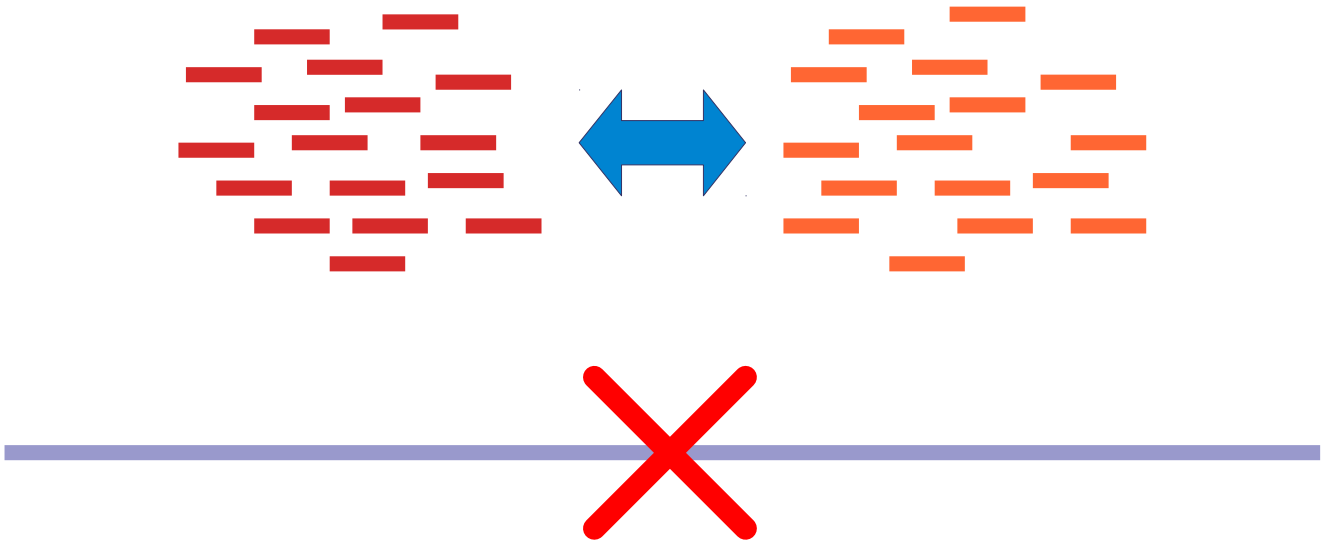
- Classical approach: mapping to a reference genome

  But

  - Not available for all species or of bad quality
  - De novo assembly is hard

Structural Variant detection

- NGS: millions of small reads



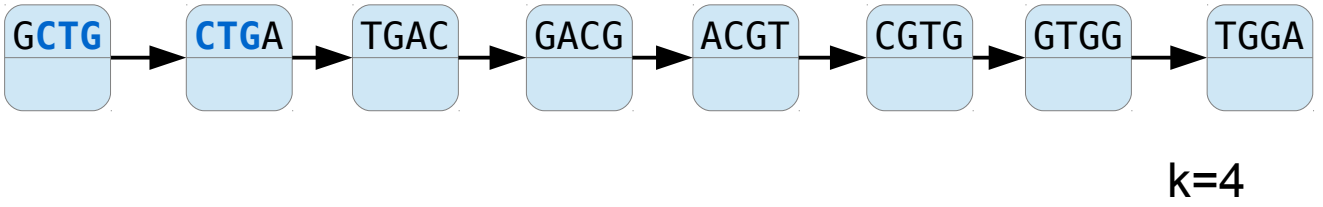
- Our approach:
  - without any reference genome
  - Without assembly

    → direct comparison of read datasets

De novo approaches

- *de Bruijn* graph
  - Definition
    - One node = one kmer
    - Edge = suffix-prefix k-1 overlaps between 2 kmers

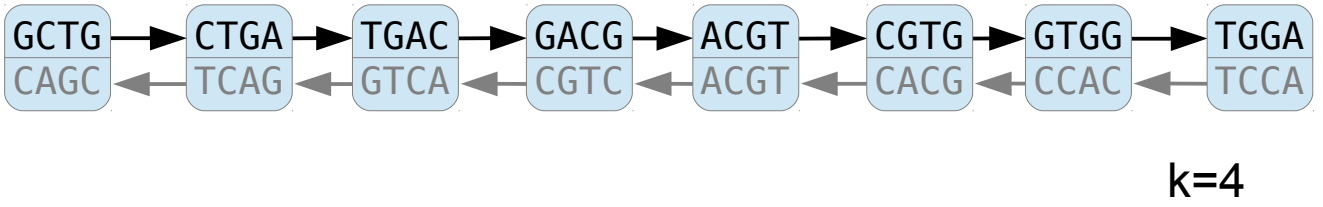
```
Seq : GCTGACGTGGA
      GCTG
       CTGA
        TGAC
         GACG
          ACGT
           CGTG
            GTGG
             TGGA
```



De novo approaches

- *de Bruijn* graph for assembling NGS reads
  - A more complex definition
    - One node = 2 kmers (forward + rev-comp)
    - Edge = idem + label for in/out strands

```
read 1  CTGACGT
read 2     ACGTGGA
read 3 CGTCAGA
```



  - Reads are cut in kmers
  - Assembly = finding long paths (> read size)

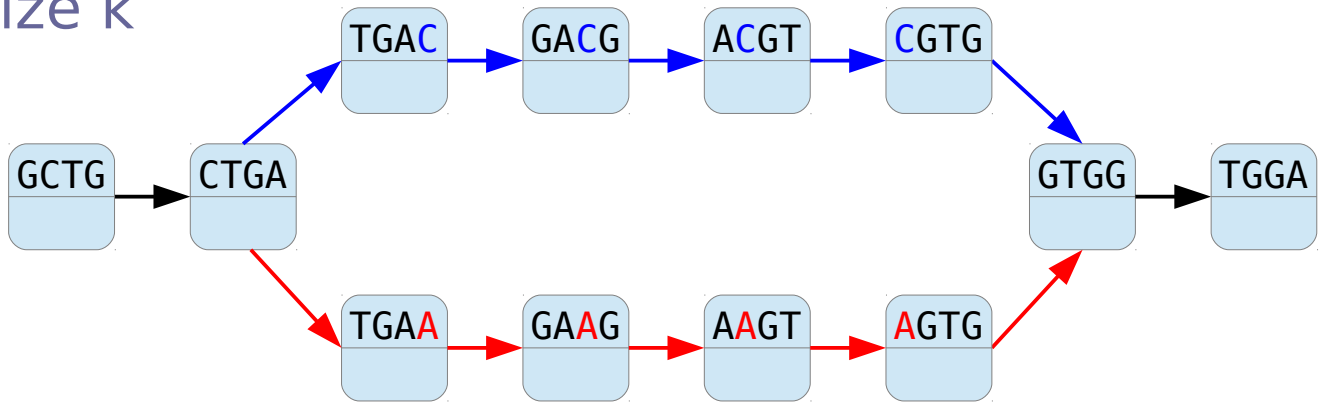
De novo approaches

- *de Bruijn* graph for variant discovery
  - Topological motifs generated by variants
  - Ex : SNPs

    Bubble of size k

Seq1 : GCTGACGTGGA

Seq2 : GCTGAAGTGGA



DiscoSnp [Uricaru *et al*. 2014]

  - Any-size bubbles: indels, splicing events (KisSplice [Sacomoto *et al.* 2012], Cortex [Iqbal *et al.* 2012])

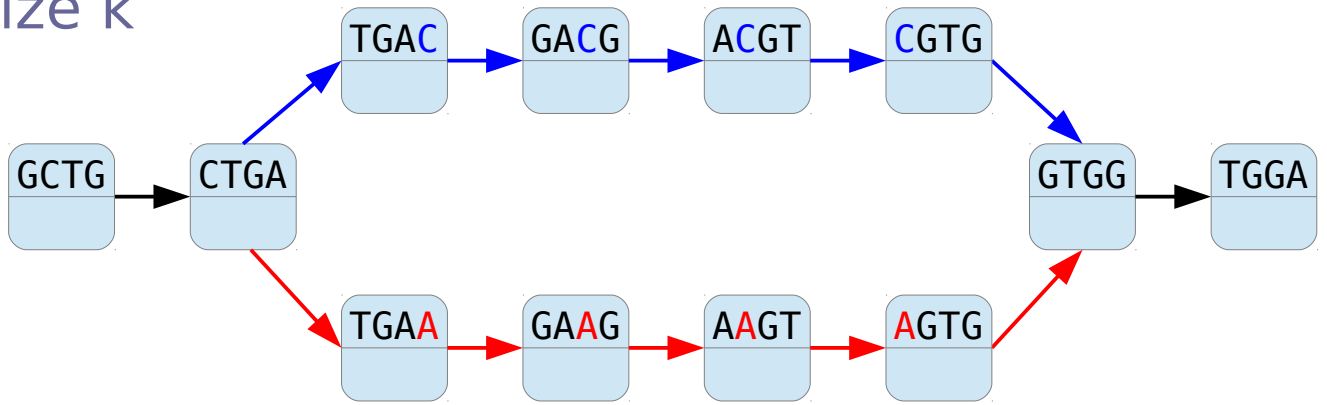
De novo approaches

- *de Bruijn* graph for variant discovery
  - Topological motifs generated by variants
  - Ex : SNPs

    Bubble of size k

Seq1 : GCTGACGTGGA

Seq2 : GCTGAAGTGGA

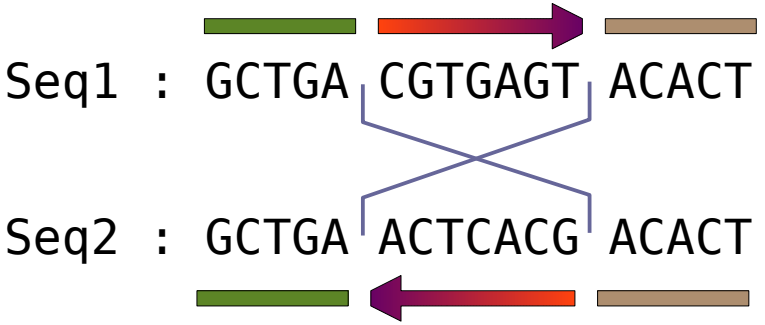


DiscoSnp [Uricaru *et al*. 2014]

  - Any-size bubbles: indels, splicing events (KisSplice [Sacomoto *et al.* 2012], Cortex [Iqbal *et al.* 2012])
  - What is the motif of inversions ?

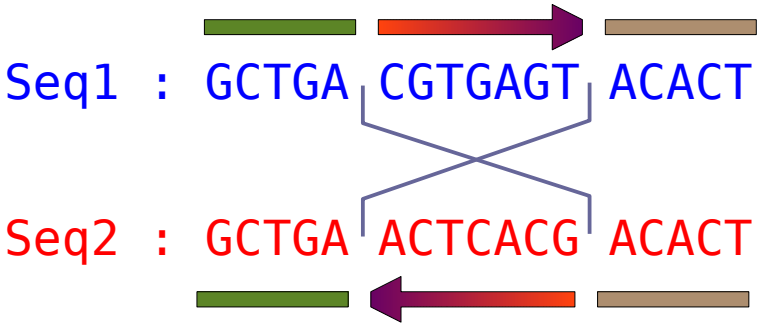
The inversion pattern

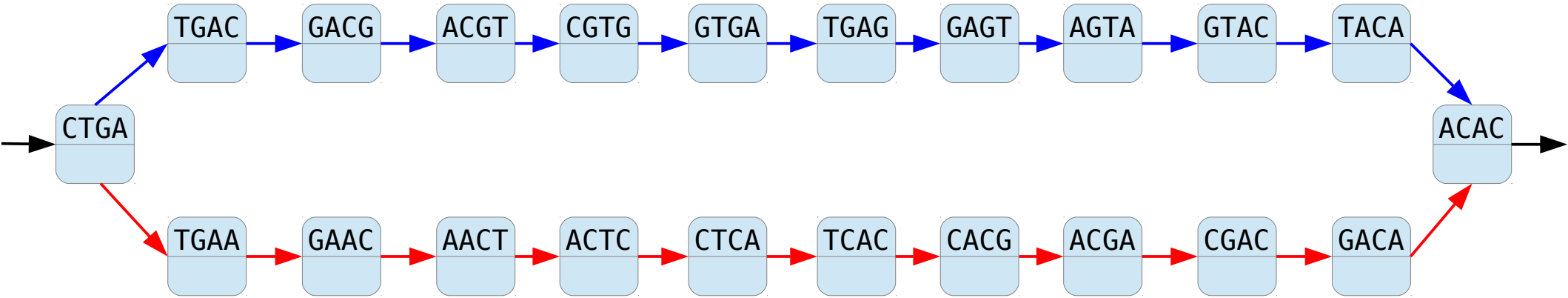
- Example :



The inversion pattern

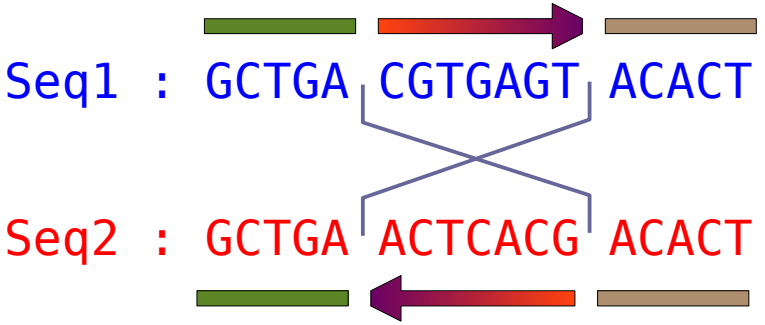
- Example :

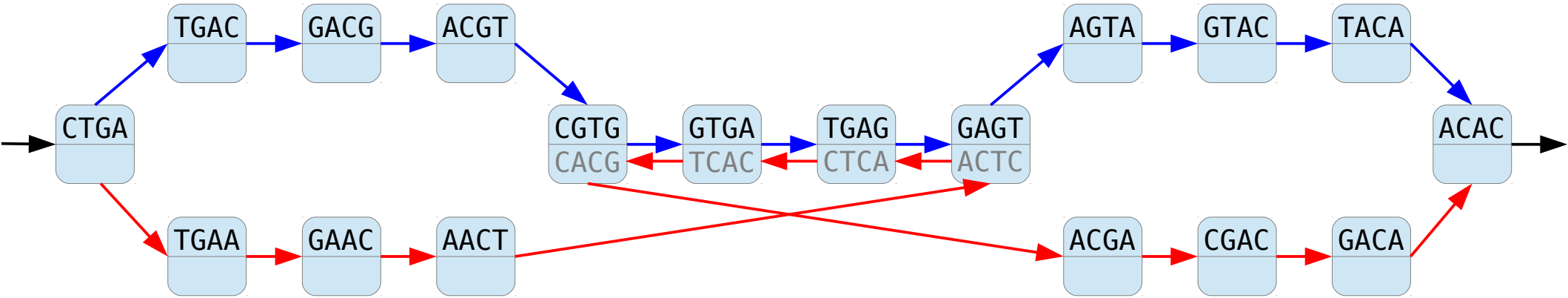




The inversion pattern

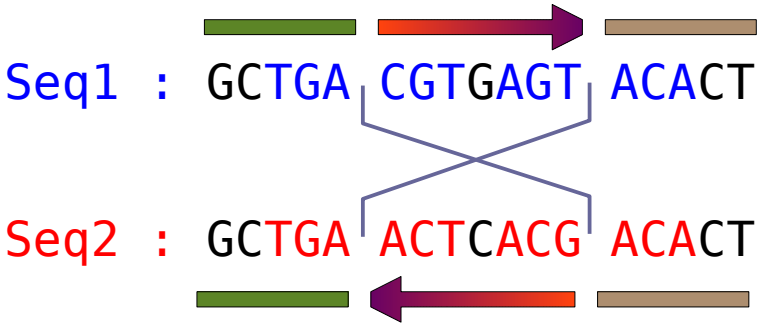
- Example :

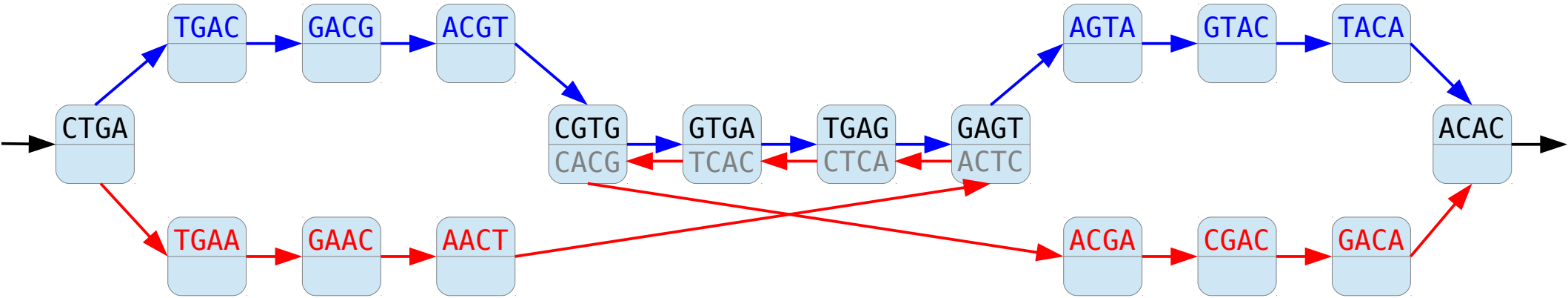




The inversion pattern

- Example :

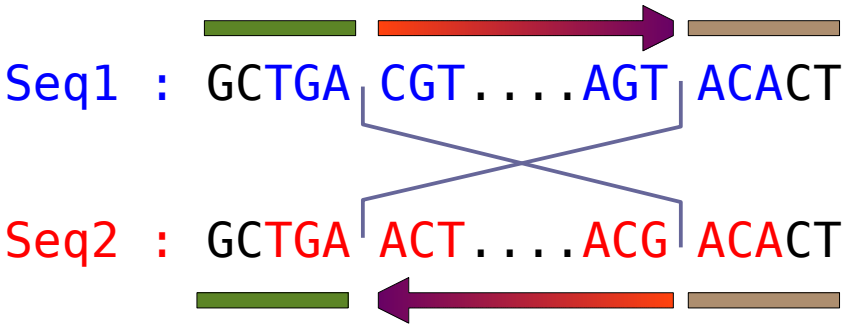


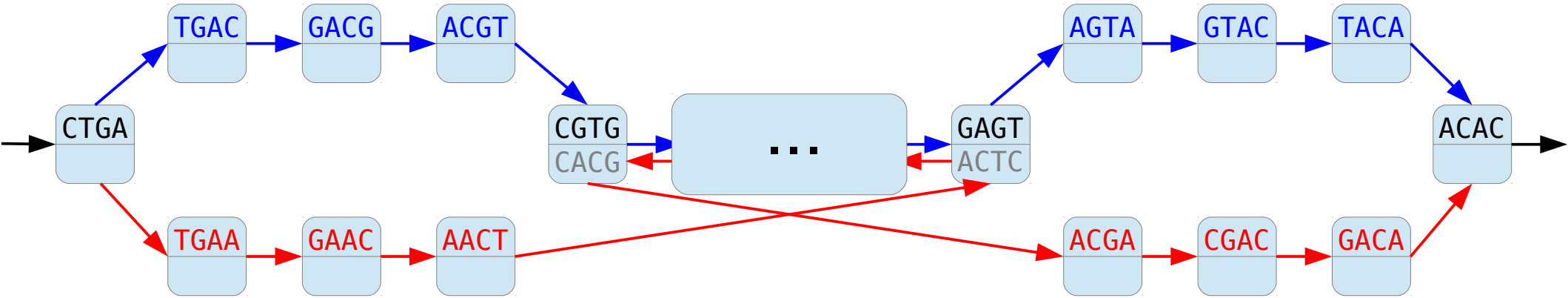


Only kmers overlap-ping the junctions are discriminant

The inversion pattern

- Example :





Only kmers overlapping the junctions are discriminant

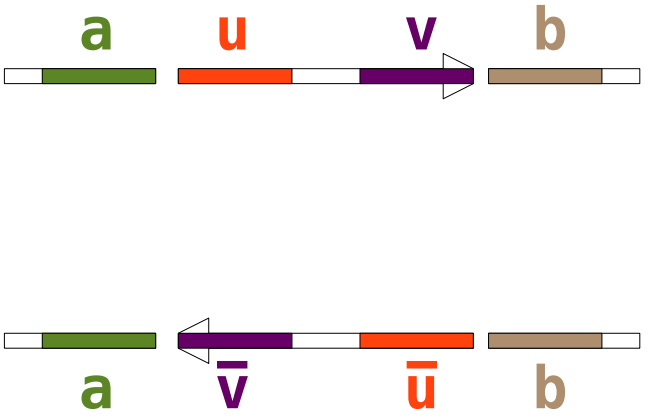
Fixed size motif independent of the inversion size

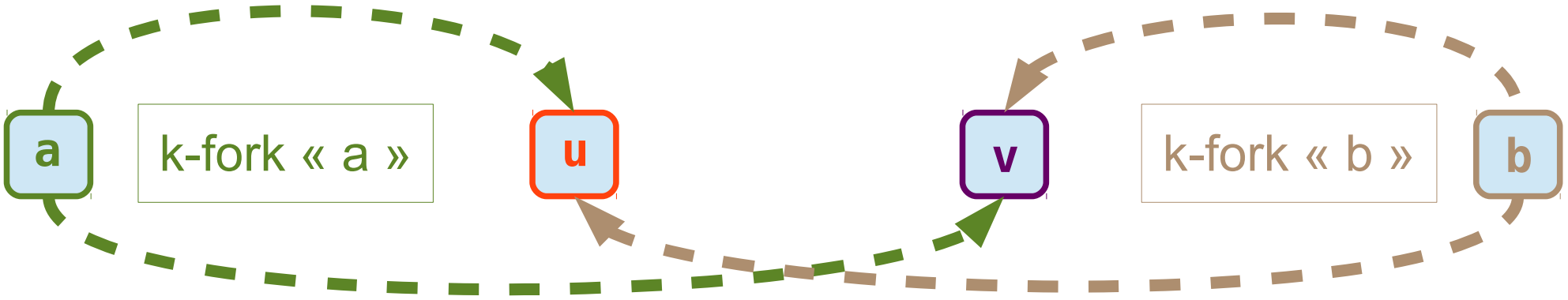
The inversion pattern

- In general:

  4 important nodes

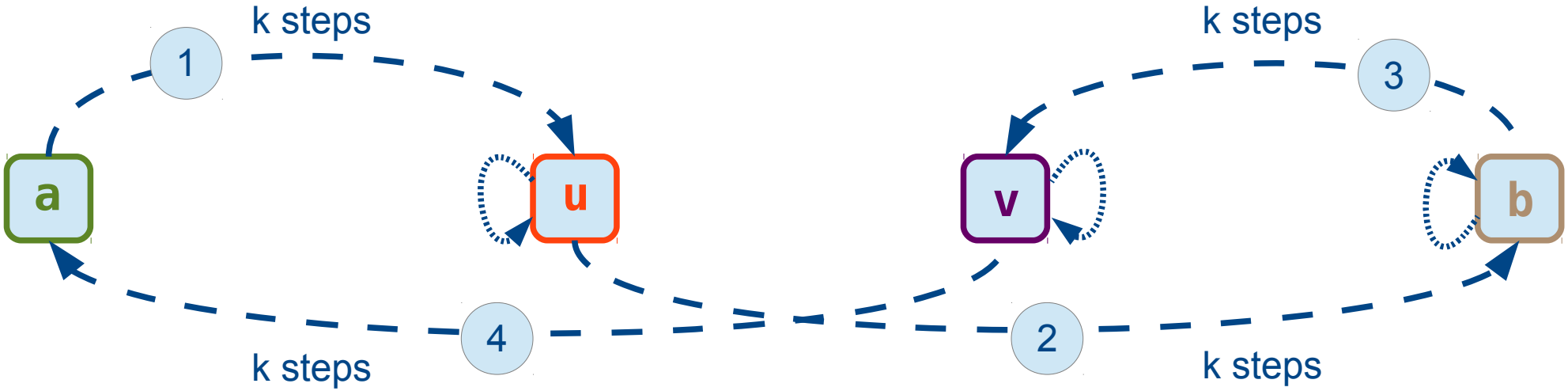
  a, **u**, **v**, **b**





Algorithm

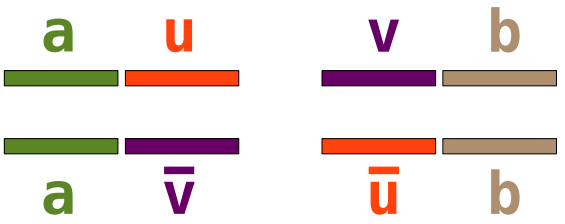
- Naïve algorithm



  - Searching 4k paths
- Improved algorithm
  - At most 2k-paths
  - Limiting search space early (see filters)

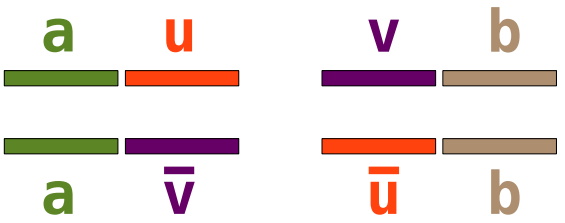
Limiting false positives

- A solution : 2 pairs of 2k-words
  - $u \neq \overline{v}$
  - $a \neq \overline{b}$



Limiting false positives

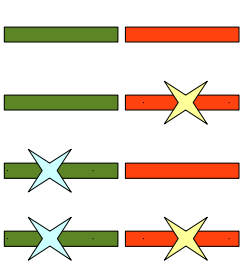
- A solution : 2 pairs of 2k-words
  - $u \neq \overline{v}$
  - $a \neq \overline{b}$



- Approximate repeats of size 2k can generate the inversion pattern

  High copy number repeats

  → combinatorial explosion



with $u \simeq \overline{v}$ and $a \simeq \overline{b}$

- 2 main filters (early during search) :
  - Similarity between sequence nodes
    $LCS(u,\overline{v}) < max\_sim$ and $LCS(a,\overline{b}) < max\_sim$
  - Local Complexity of the graph
    # of k-neighbors < LCT

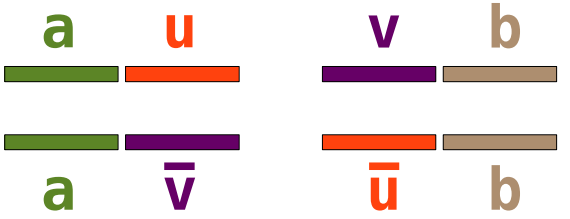
Implementation

- Light de Bruijn Graph representation :
  - From minia assembler [Chikhi and Rizk, 2012]
  - Bloom filters, <5GB for a human genome
  - GATB C++ library [Drezen et al. 2014] http://gatb.inria.fr/
- Software **TakeABreak**

  http://colibread.inria.fr/fr/software/takeabreak/
  - Input : n (1 → N) read datasets (fasta, fastq, gz)
  - Parameters :

    De Bruijn graph: k, frequency threshold

    Inversions: max_sim, LCT
  - Output : pairs of breakpoint sequences



Evaluation on simulated datasets

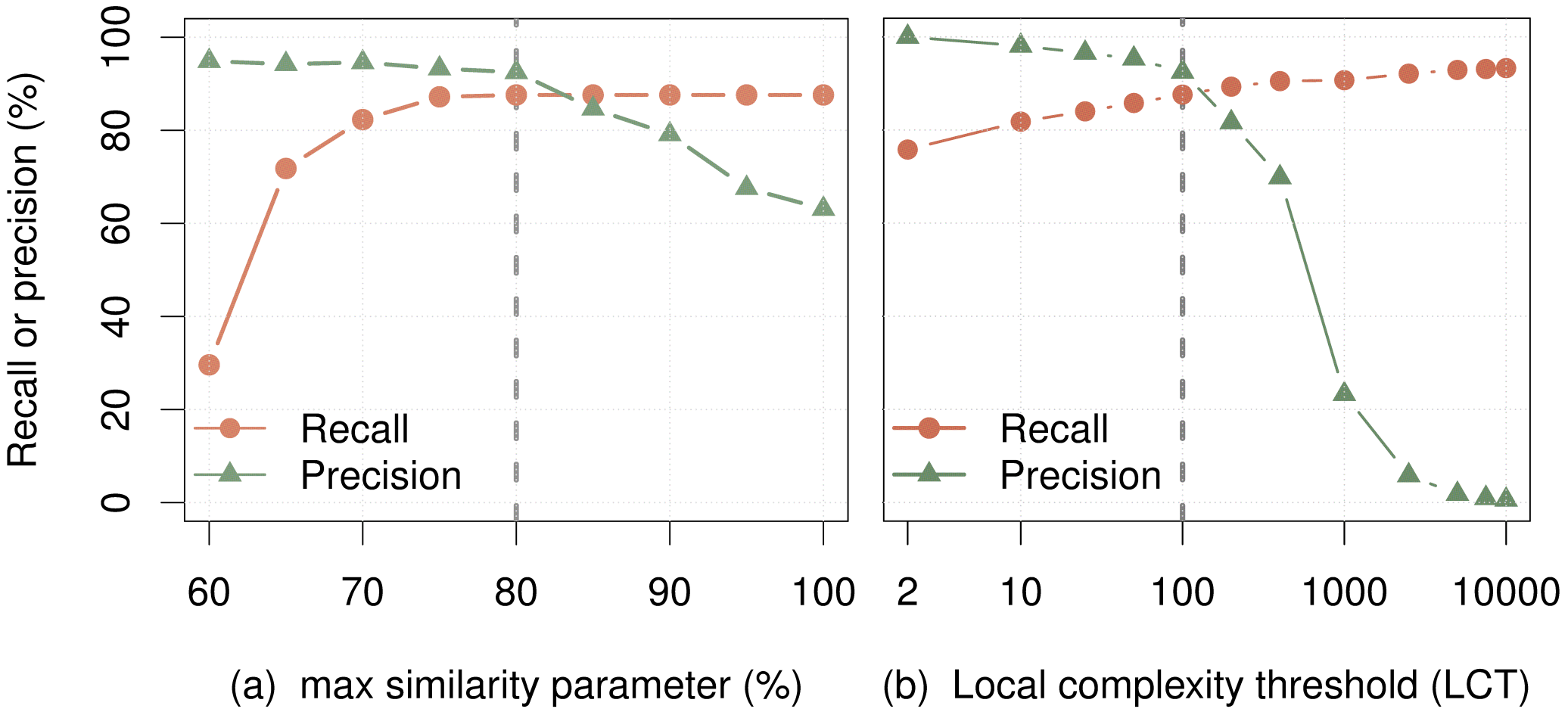
- Simulations:
  - Real genome : *E. coli* (5 Mbp), *C. elegans* (100 Mbp), human chr 22 (35 Mbp)
  - Simulate 1000 non-overlapping inversions
  - Simulate 40x 100 bp reads on each genome (1 % error)
- Results: good recall and precision

	Recall (%)	Precision (%)	# FP
E. coli	100.0	100.0	0
C. elegans	96.0	99.1	9
Human chr 22	87.6	92.5	71

Default parameters : k=31, freq =3, max_sim = 80 %, LCT = 100

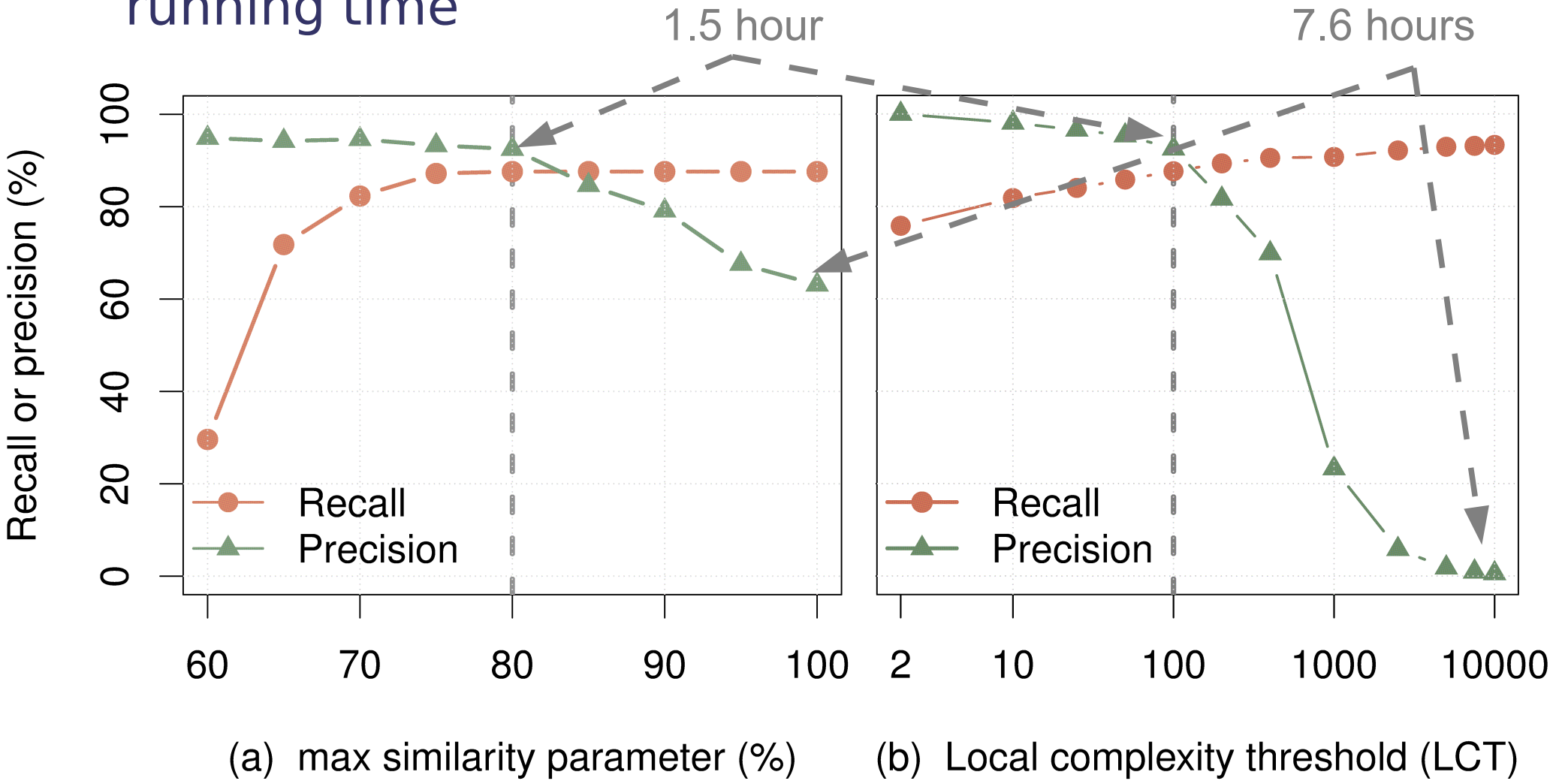
Effect of the parameters

on Human chr 22 : limiting false positives



Effect of the parameters

on Human chr 22 : limiting false positives and running time



Comparisons

- Other SV discovery tools :
  - Without a reference genome : Cortex [Iqbal et al. 2012]

    0 % recall for inversions : only « clean » bubbles
  - With a reference genome : Breakdancer [Chen et al. 2009]

    on human Chr 22 dataset:

    1200 inversion calls for 1000 simulated inversions

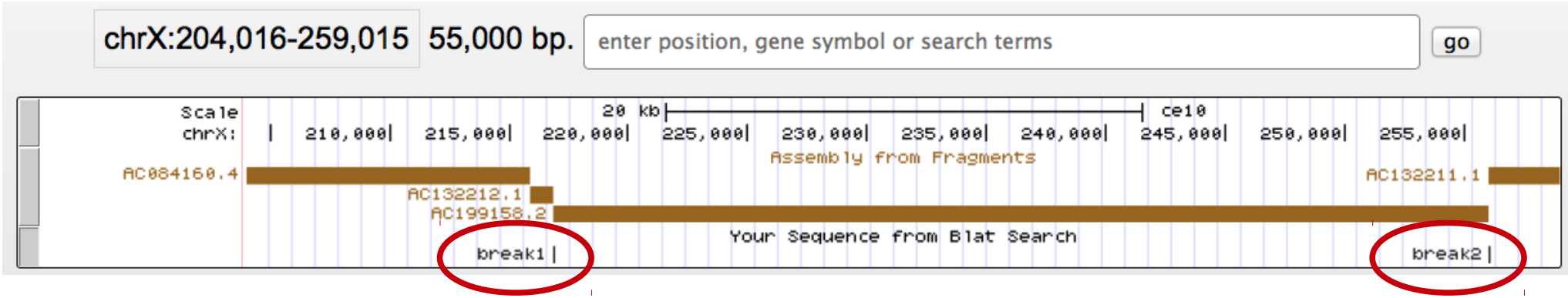
    imprecise coordinates

    (here requiring >50 % overlap of inverted segment)

	Recall (%)	Precision (%)
Breakdancer	81.4	69.8
TakeABreak	87.6	92.5

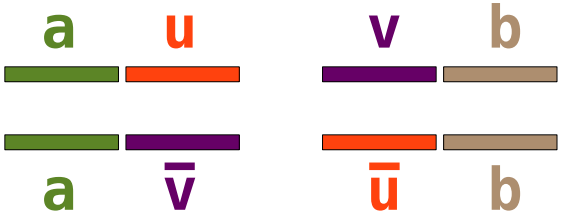
Real data

- Data :
  - *C. elegans* SRR065390 read dataset
    66 M 100bp Illumina reads ~ 70x coverage
  - Simulate 1000 inversions in reference genome and 70x read dataset on mutated genome
- Results :
  - Calls : 991, with at least 956 « true » inversions
  - <1.5 hour on a laptop (<2 GB memory – 14 Gbp)
  - A putative scaffolding error in the reference genome?



Real inversions ?

- Still looking for a reference set of validated inversions...
- Known limitations :
  - « Clean » breakpoints



  - No inverted repeat ≥ k at the breakpoints ($u \neq \overline{v}$ and $a \neq \overline{b}$)
- Real inversions
  - SNPs/indel inside the breakpoints
  - Known mechanisms (NHEJ, NAHR) → small indels or inverted repeats

TakeABreak – recap

- Inversion breakpoint discovery
  - The only method without any reference genome
  - Easy to use, few parameters
  - Fast and low memory
  - Proof of concept on simulated inversions
- Future work:
  - More flexible motif
  - Assembling the inverted segment

    MindTheGap : fill the gap between u → v

Thanks !

Pierre Peterlongo

Liviu Ciortuz
now in Romania, Iasi

Genscale team in Rennes (France)

https://team.inria.fr/genscale/

Dominique Lavenier Erwan Drezen Guillaume Rizk

Colib'read

ANR Colib'read

http://colibread.inria.fr/

Région Bretagne

ANR GATB

http://gatb.inria.fr/

References

- DiscoSnp. Reference-free detection of isolated SNPs. R. Uricaru, G. Rizk, V. Lacroix, E. Quillery, O. Plantard, R. Chikhi, C. Lemaitre, P. Peterlongo. *Under review*
- Cortex. De novo assembly and genotyping of variants using colored de Bruijn graphs. Z. Iqbal, M. Caccamo ,I. Turner, P. Flicek, G. McVean. *Nature Genetics*, 2012, 44, 226-–232
- Minia. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. R. Chikhi, G. Rizk. *WABI* 2012, 7534, 236-248
- GATB. GATB: Genome Assembly & Analysis Tool Box. E. Drezen, G. Rizk, R. Chikhi, C. Deltel, C. Lemaitre, P. Peterlongo, D. Lavenier. To appear in *Bioinformatics*
- Breakdancer. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Chen, *et al*. *Nat Method*, 2009, 6, 677-681