reads2genpop: From sequence reads to genomes and populations

September 21 – 22, 2022

Methodological challenges of Structural Variation characterization and the particular case of insertions

Claire Lemaitre

Genscale, Inria Rennes Bretagne Atlantique – IRISA, Rennes

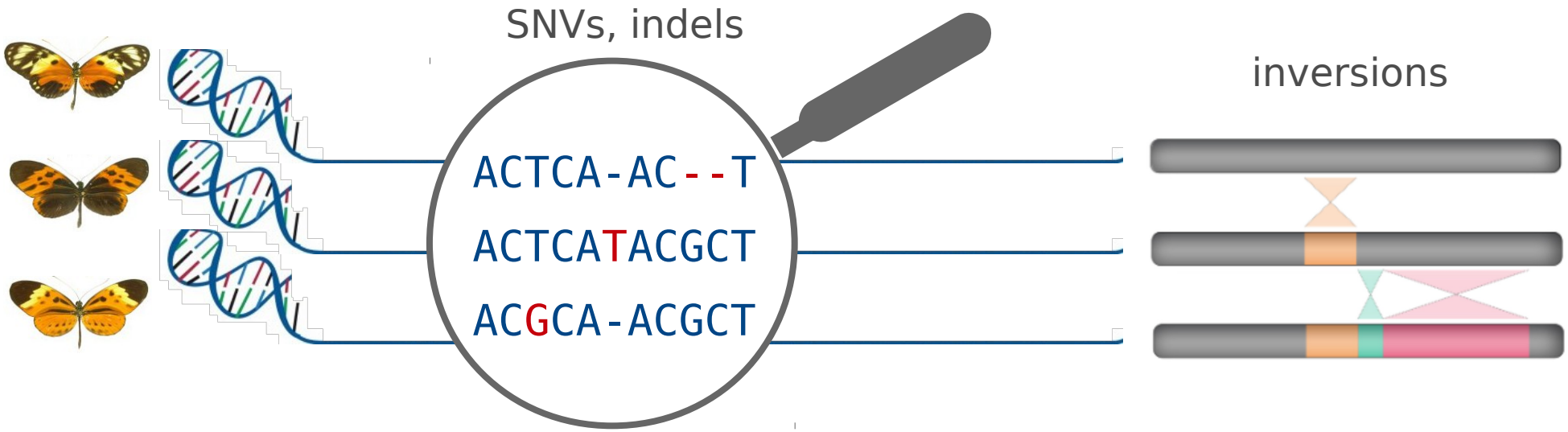
claire.lemaitre@inria.fr

http://people.rennes.inria.fr/Claire.Lemaitre/

Genetic variations

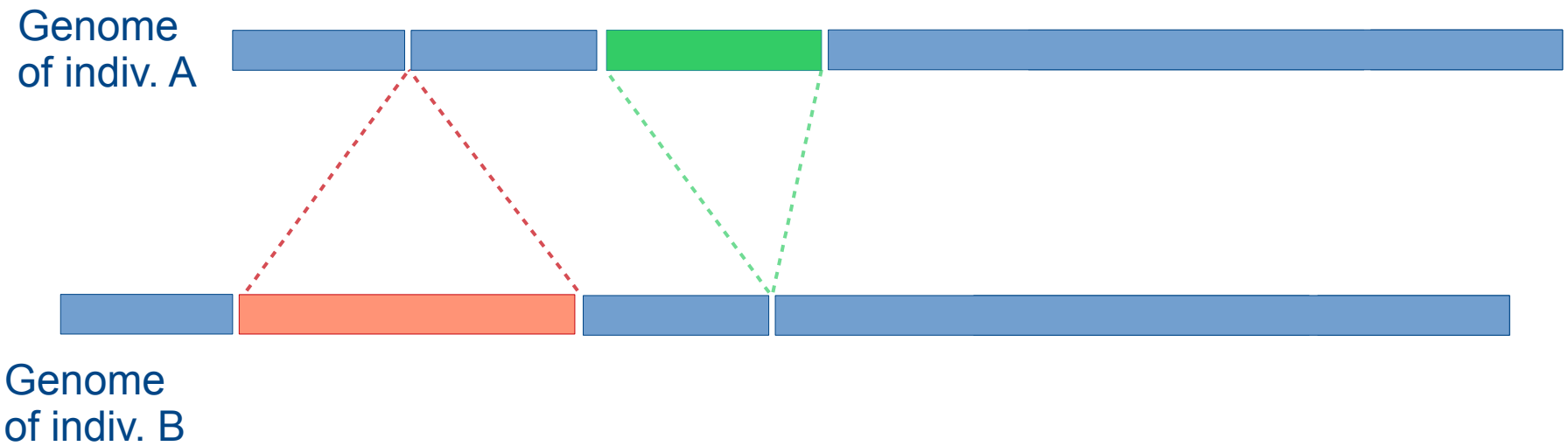
- Intra-species diversity

- Genomic variants : from punctual to large differences between the genomes



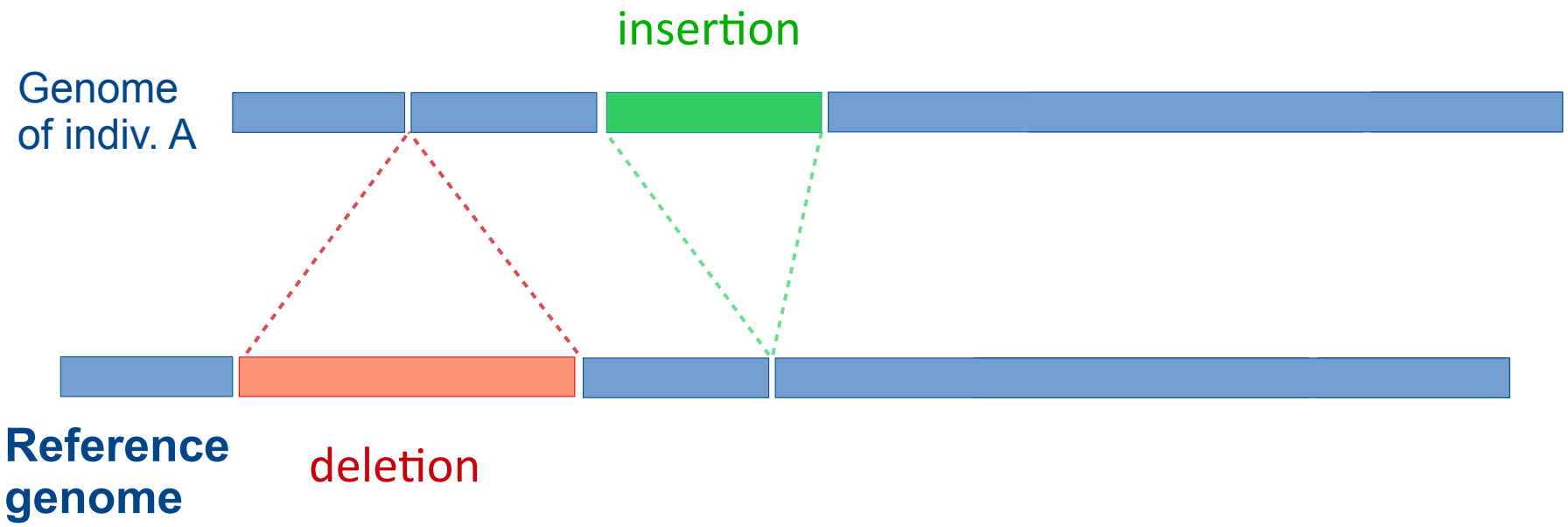
Structural Variants

- A simple definition : genome variation of size > 50 bp
- That gathers many different types



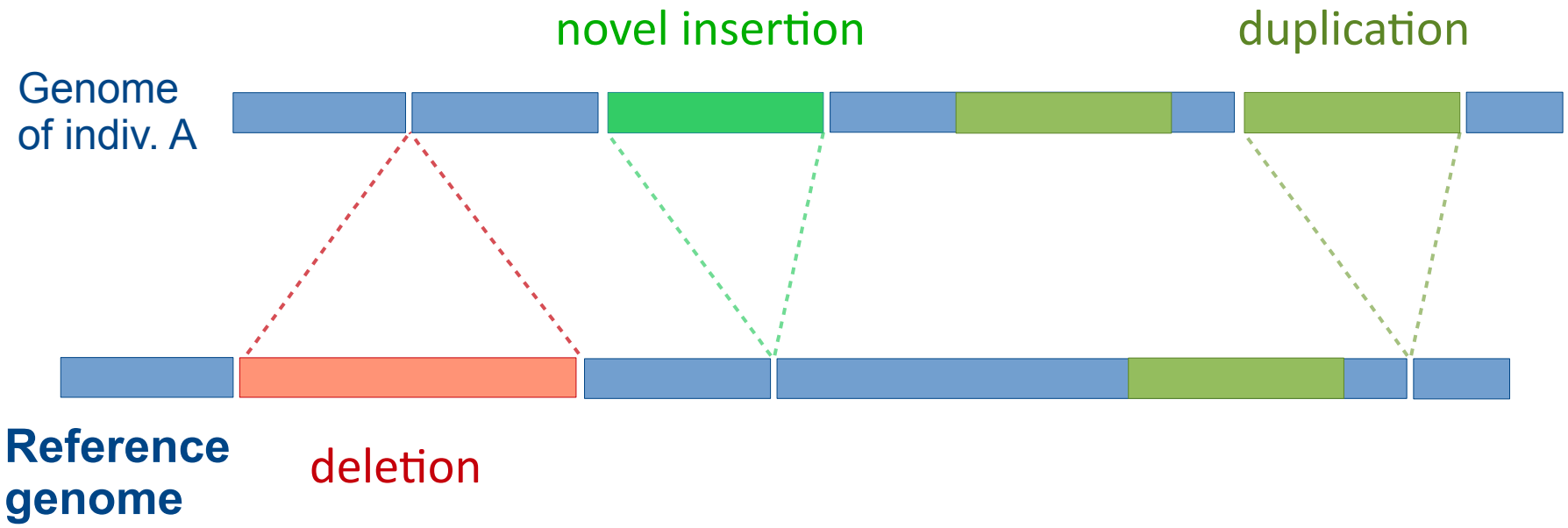
Structural Variants

- A simple definition : genome variation of size > 50 bp
- That gathers many different types



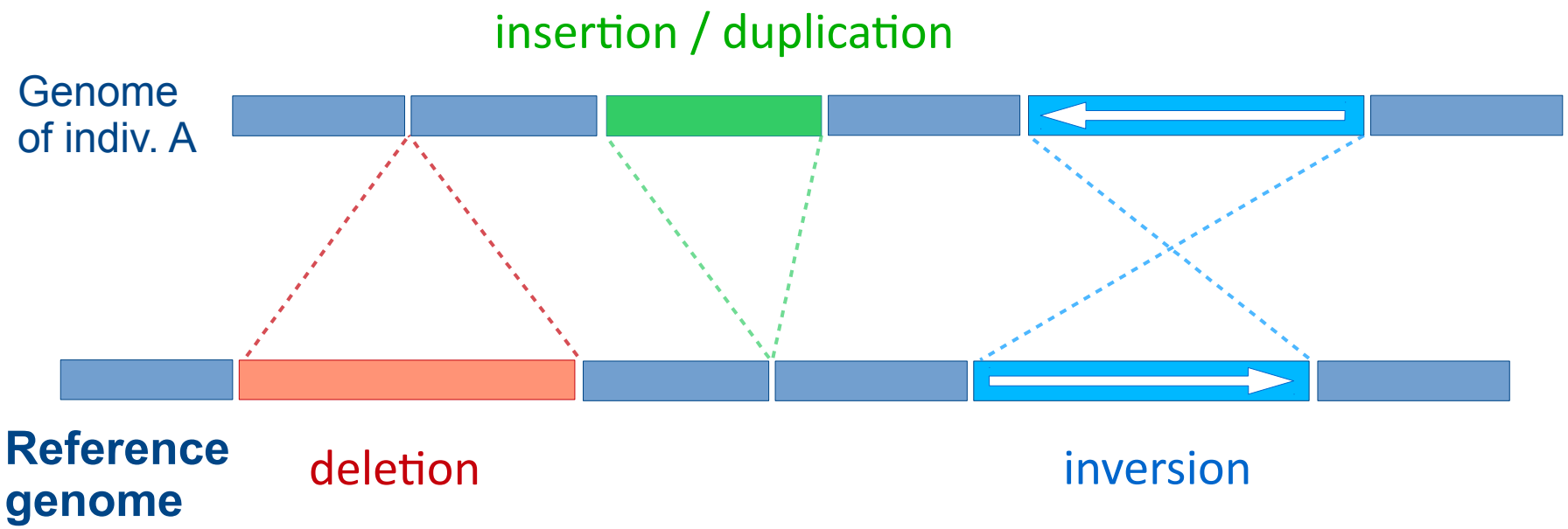
Structural Variants

- A simple definition : genome variation of size > 50 bp
- That gathers many different types



Structural Variants

- A simple definition : genome variation of size > 50 bp
- That gathers many different types



and transpositions, translocations...

Why studying SVs ?

- SV events are 10 to 100 times less frequent than SNPs…

  … but involve 15 times more base pairs [in humans: Pang *et al*, 2010]

- Various functional and evolutionary impacts

  modifying functional elements, expression levels, suppressing recombination, such as :

  - human diseases : repeat expansion in Parkinson, gene fusions in cancer
  - plant selection : [Alonge et al, 2020]

    100 tomato varieties : 240,000 SVs
    (causal) association with flavor, size and yield

  - evolution : inversions and supergenes

In this talk

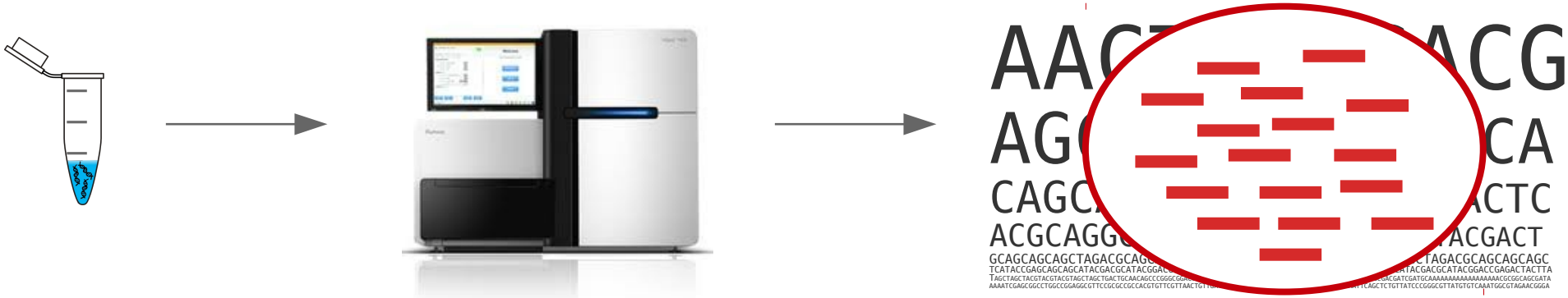
- How do we discover SVs with sequencing data ?
  - Overview of the different approaches
  - Main problems/challenges
  - Short State of the Art
- The case of long insertions :
  - Very difficult type with short reads
  - Thanks to long reads, analysing real insertion variants & revisiting short-read based results
- After the discovery...

In this talk

1. How do we discover SVs with sequencing data ?
   - Overview of the different approaches
   - Main problems/challenges
   - Short State of the Art

Accessing the genomes of many individuals

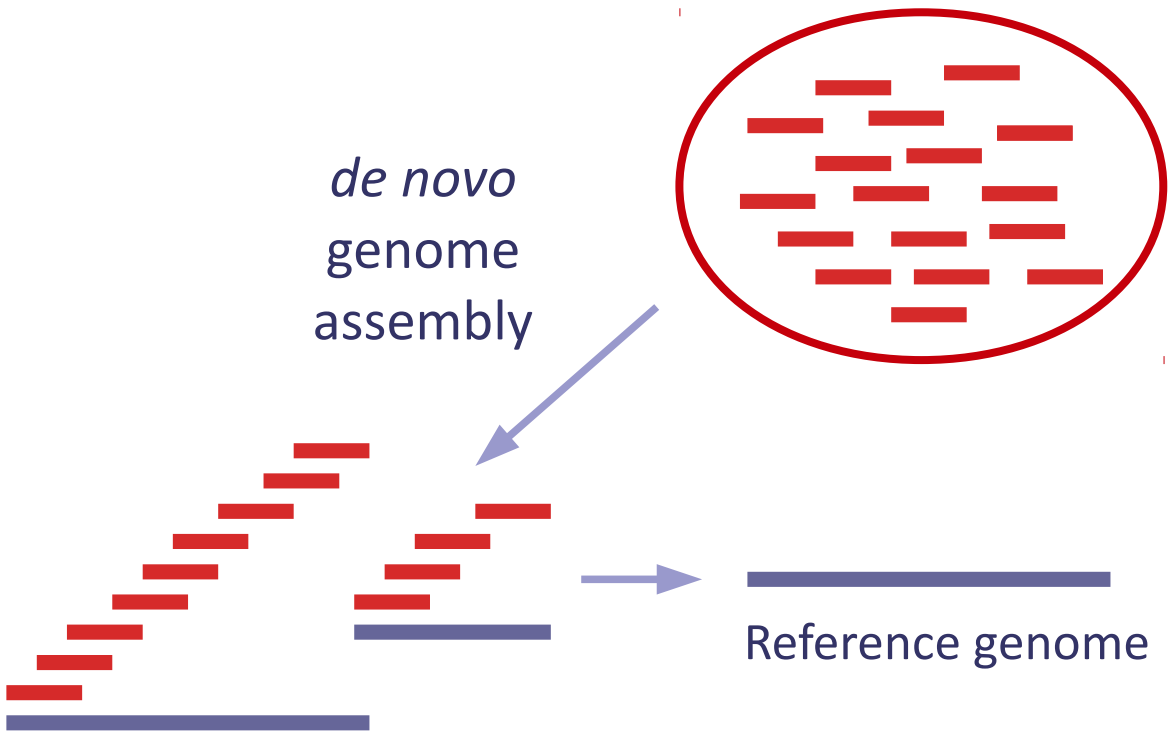
- Since 2008, high throughput sequencing :



- Data = sets of many small sequences (*reads*)
- 2 types of sequencing data :
  - Short reads (2008...) : ~ 100 bp, <0.1 % error rate
  - Long reads (2015...) : 1 – 1,000 Kb, 5 – 15 % error rate

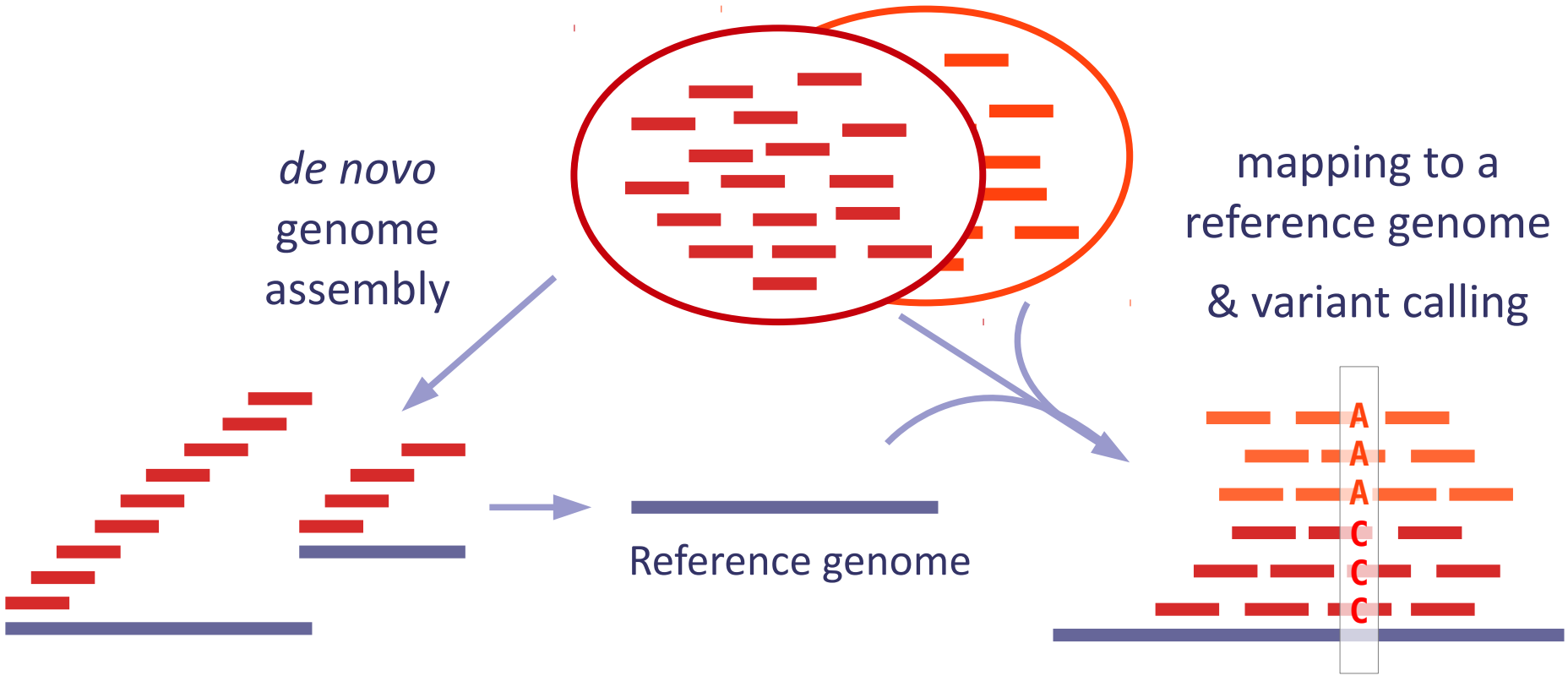
Methods for sequencing data

- Two classical approaches :
  - Sequence assembly : hard problem, resource-consuming



Bioinformatics

- Two classical approaches :
  - Sequence assembly : hard problem, resource-consuming
  - Read mapping & variant calling : relying on a reference genome



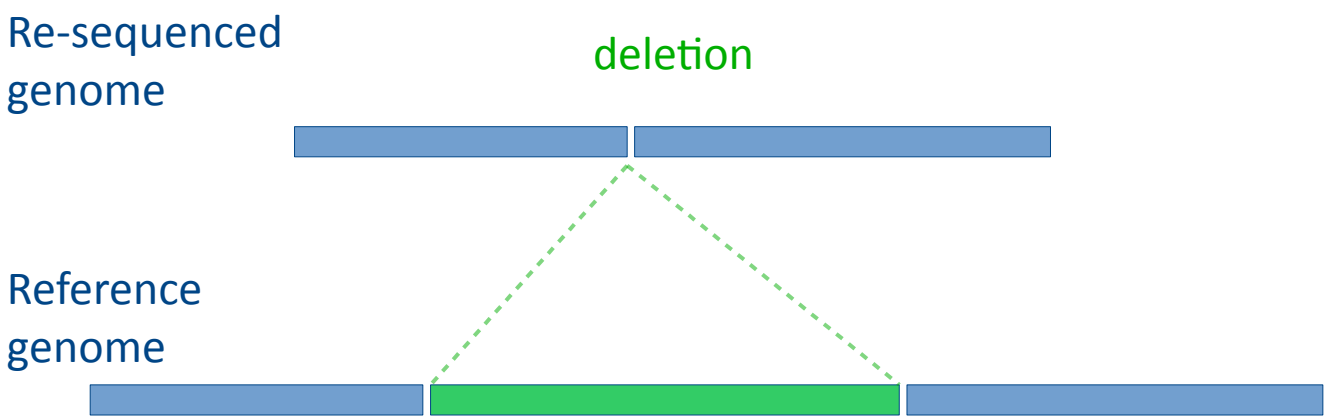
Calling Structural Variants

- A much more difficult problem than SNV-indel calling :

  The whole alternative allele is not found in a single read alignment

  → Looking for aberrant combinations of several alignments

- 3 types of mapping signals :



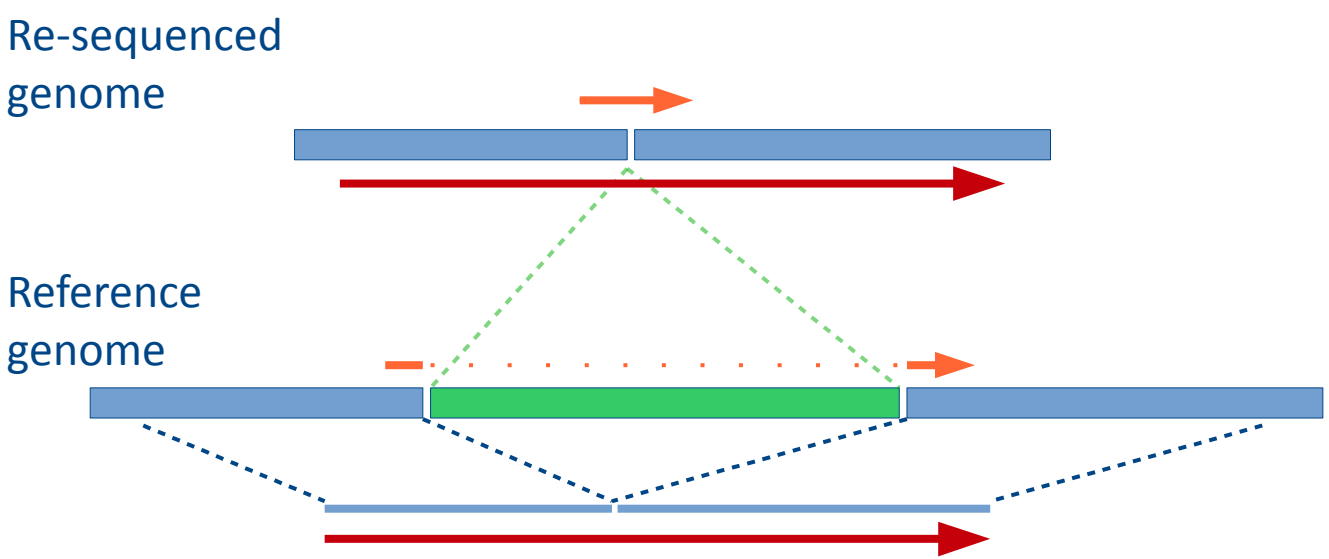
Calling Structural Variants

- A much more difficult problem than SNP-indel calling :

  The whole alternative allele is not found in a single read alignment

  → Looking for aberrant combinations of several alignments

- 3 types of mapping signals :
  - Split-read



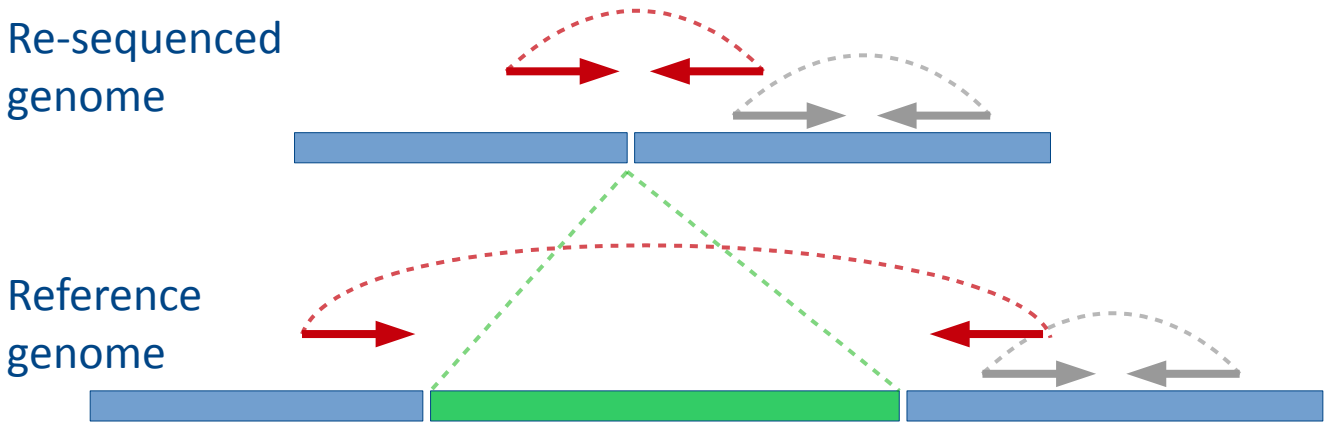
Calling Structural Variants

- A much more difficult problem than SNP-indel calling :

  The whole alternative allele is not found in a single read alignment

  → Looking for aberrant combinations of several alignments

- 3 types of mapping signals :
  - Split-read
  - Paired mapping



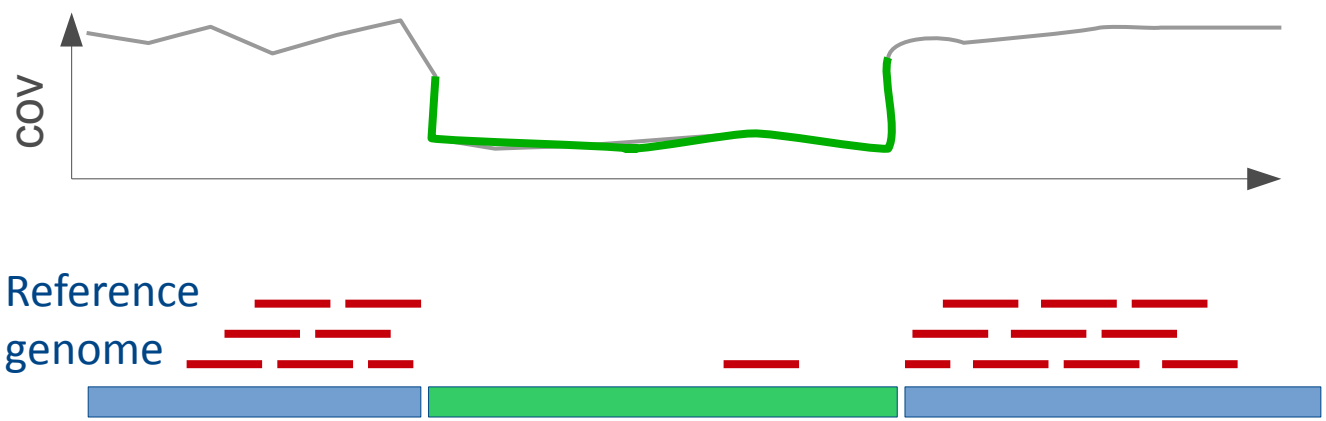
Calling Structural Variants

- A much more difficult problem than SNP-indel calling :

  The whole alternative allele is not found in a single read alignment

  → Looking for aberrant combinations of several alignments

- 3 types of mapping signals :
  - Split-read
  - Paired mapping
  - Read depth



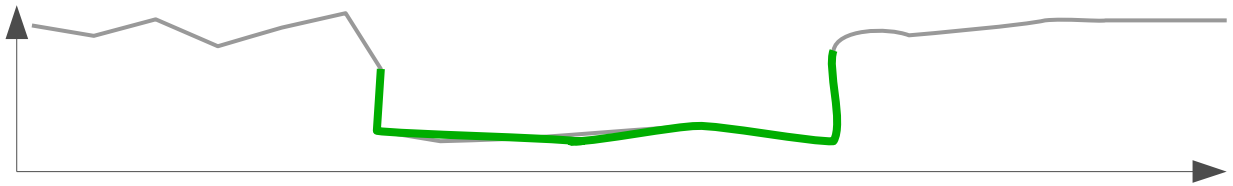
Calling Structural Variants

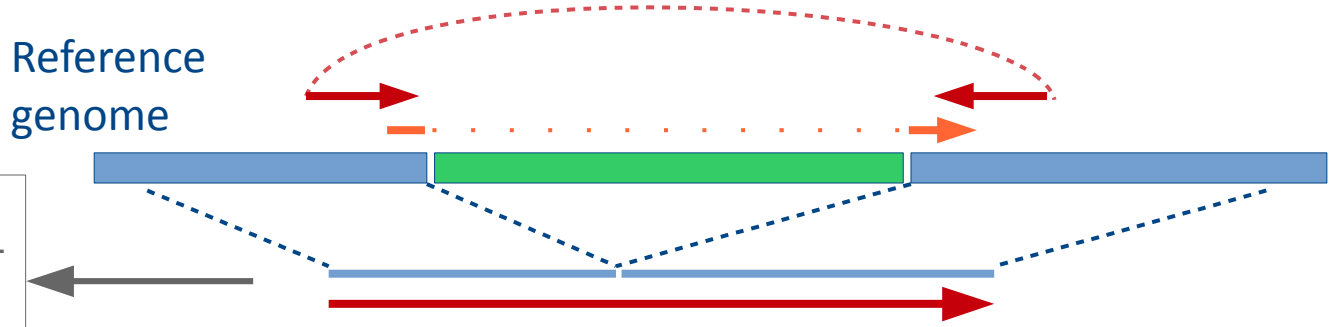
- A much more difficult problem than SNP-indel calling :

  The whole alternative allele is not found in a single read alignment

  → Looking for aberrant combinations of several alignments

- 3 types of mapping signals :
  - Split-read
  - Paired mapping
  - Read depth

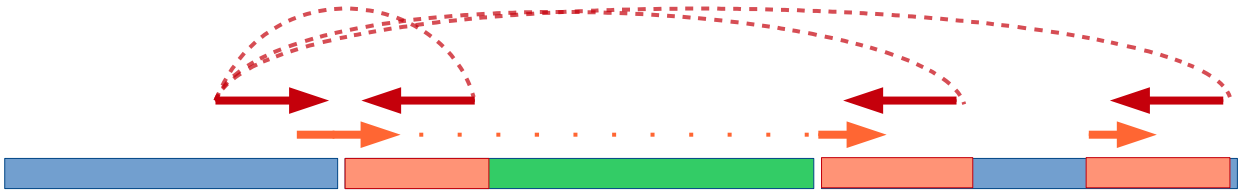





```
VCF :
#chrom   pos   REF       ALT
chr3     1562  ACCTATG   A
```

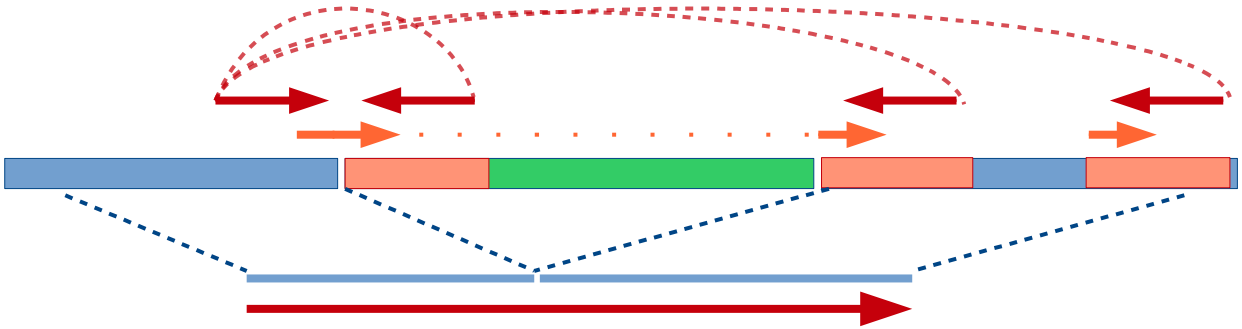
SV calling : read size matters

- Difficulties :
  - Heterogeneity of types → no equivalence 1 signal ↔ 1 type
  - Genome repeats
    - Mapping ambiguities → False Positive calls
    - SVs are associated with repeats → Missing calls (False Negatives)



SV calling : read size matters

- Difficulties :
  - Heterogeneity of types → no equivalence 1 signal ↔ 1 type
  - Genome repeats
    - Mapping ambiguities → False Positive calls
    - SVs are associated with repeats → Missing calls (False Negatives)



Advantages of long reads : can contain the alternative allele and span the repeats

History of the art

- 2008 – 2018 : more than 70 SV callers for short reads
  - At first, 1 signal at a time
    ex : BreakDancer (Chen 2009), Pindel (Ye 2009), CNVnator (Abyzov 2011)
  - Then, combining several signals
    ex : Delly (Rausch 2012), Lumpy (Layer 2014)…
  - Some « meta » SV callers :
    ex : metaSV (Mohiuddin 2015), Parliament (English 2015)
  - Last generation, use of assembly techniques
    ex : Manta (Chen 2016), GRIDSS (Cameron 2017), Svaba (Wala 2018)

Some reviews : (Medvedev *et al*, Nat Met 2009) (Alkan *et al*, Nat Rev Genet 2011)

History of the art (2)

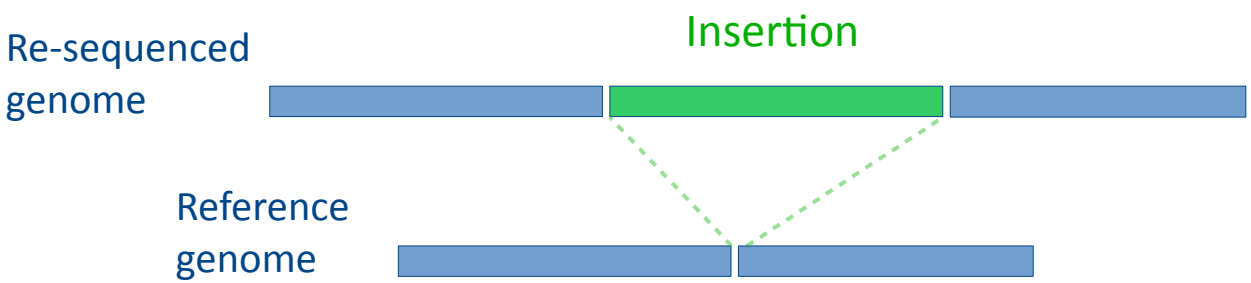
- 2008 – 2018 : more than 70 SV callers for short reads
  - Poor results : small overlap between tools
  - Benchmarks : very few and late (Kosugi 2019, Cameron 2019)
    Results : low recall (10-70%), high FP rate (up to 90 %)
  - Applications restricted to deletions (ex: 1000 genome project)

History of the art (2)

- 2008 – 2018 : more than 70 SV callers for short reads
  - Poor results : small overlap between tools
  - Benchmarks : very few and late (Kosugi 2019, Cameron 2019)
    Results : low recall (10-70%), high FP rate (up to 90 %)
  - Applications restricted to deletions (ex: 1000 genome project)
- 2018 – now : long reads = a big change for SV analysis
  - Efficient tools, based on split-mapping (main issue = mapping)
    ex : Sniffles (Sedlazeck 2018), Pbsv (Pacific Biosciences), SVIM (Heller 2019)
  - High quality SV data for applications and benchmarking
- More recent reviews : (Ho *et al*, Nat Rev Genet 2019) (Mahmoud *et al*, Genome Biol 2019)

In this talk

1. How do we discover SVs with sequencing data ?
2. The case of long insertions

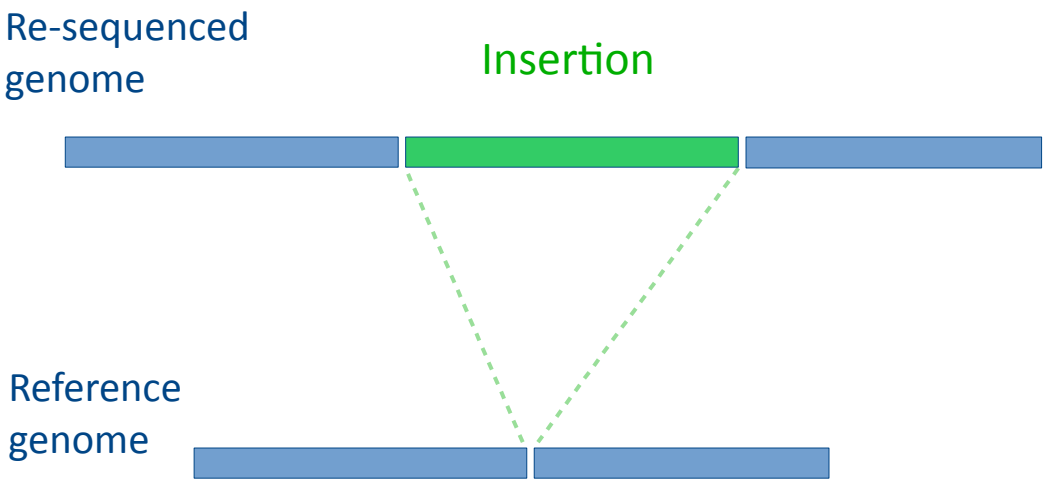


- Very difficult type with short reads
- Thanks to long reads, analysing real insertion variants & revisiting short-read based results

Insertion variants : a most difficult type of SV

- Insertions variants :
  - As frequent as deletions (inverse event)
  - But under-represented in databases :

    In dbVar : 28 % vs 72 % (deletions) – only 1.5 % with sequence resolution

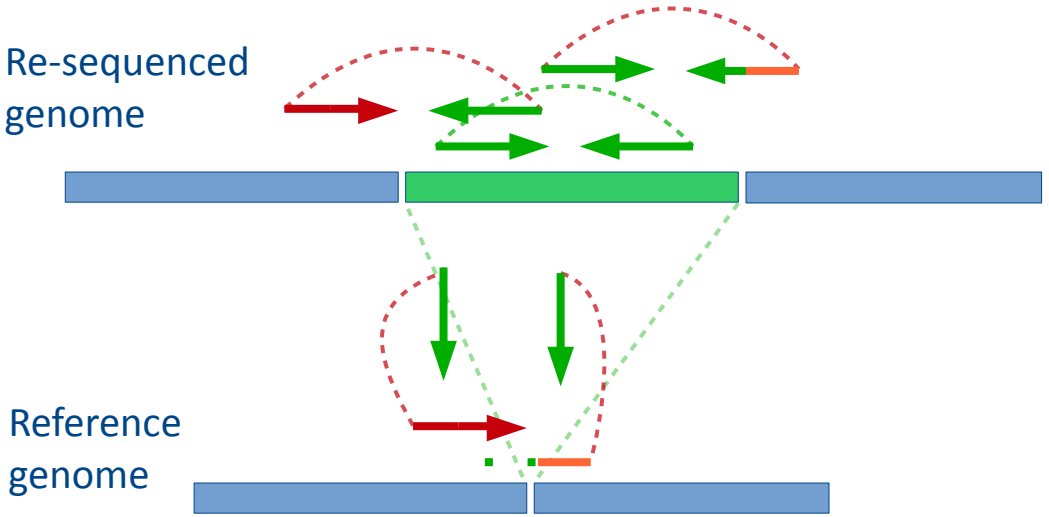


Insertion variants : a most difficult type of SV

- Insertions variants :
  - As frequent as deletions (inverse event)
  - But under-represented in databases :

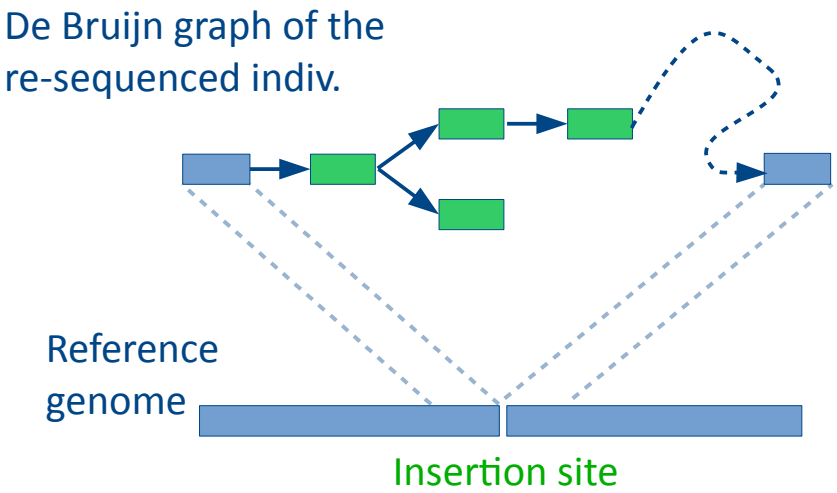
    In dbVar : 28 % vs 72 % (deletions) – only 1.5 % with *sequence resolution*

- 2 problems in 1 :
  - Insertion site :
    less mapping signals
  - Inserted sequence :
    unmapped (or far away) reads



De novo assembly for long insertion calling

- Inserted sequence recovery : need of *de novo* assembly with short reads
- MindTheGap:
  - Detection and *de novo* assembly of inserted sequences with a de Bruijn Graph
  - First tool using the whole read set
  - Results :
    - No competitor for long (>100 bp) insertions (in 2014...)
    - Very good results on simulated data...



Rizk *et al*, *Bioinformatics*, 2014.

https://github.com/GATB/MindTheGap

BIOCONDA

New insights with long read technologies

- 2 major papers in 2019-2020 : HGSV and Genome in a Bottle consortiums [Chaisson et al, 2019 and Zook et al, 2020]
  - >10 sequencing tech. and many assembly & SV calling software
  - **gold standard** SV callsets for 4 human individuals

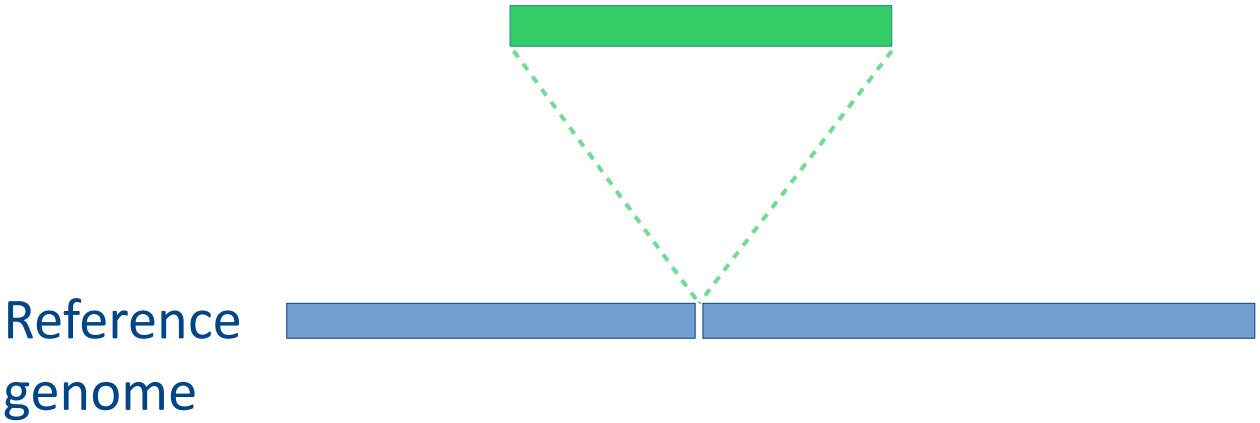
    ~ 30,000 SVs per indiv. : 50 % deletions / 50 % insertions

    all sequenced-resolved
- Bad surprise for short-read insertion callers : very low recall for MindTheGap (and other tools) : 2-10%

What make those variants so difficult to be discovered (vs simulated ones) ?

Fine characterization of real human insertions

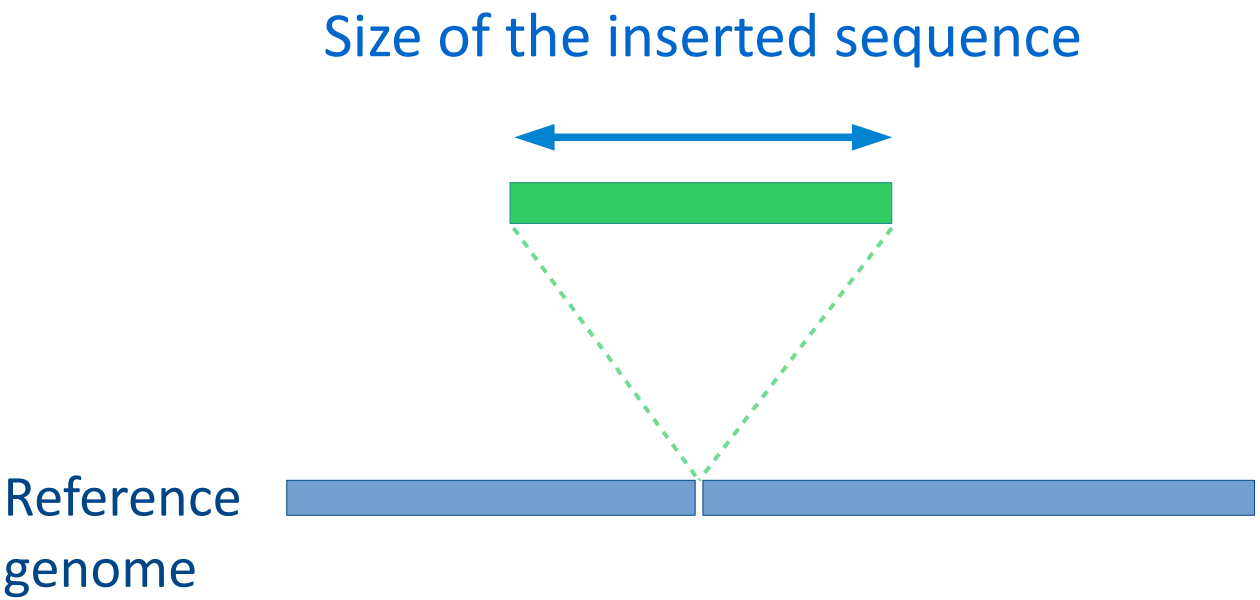
– Method : 4 levels of characterization



Fine characterization of real human insertions

– Method : 4 levels of characterization

1. size



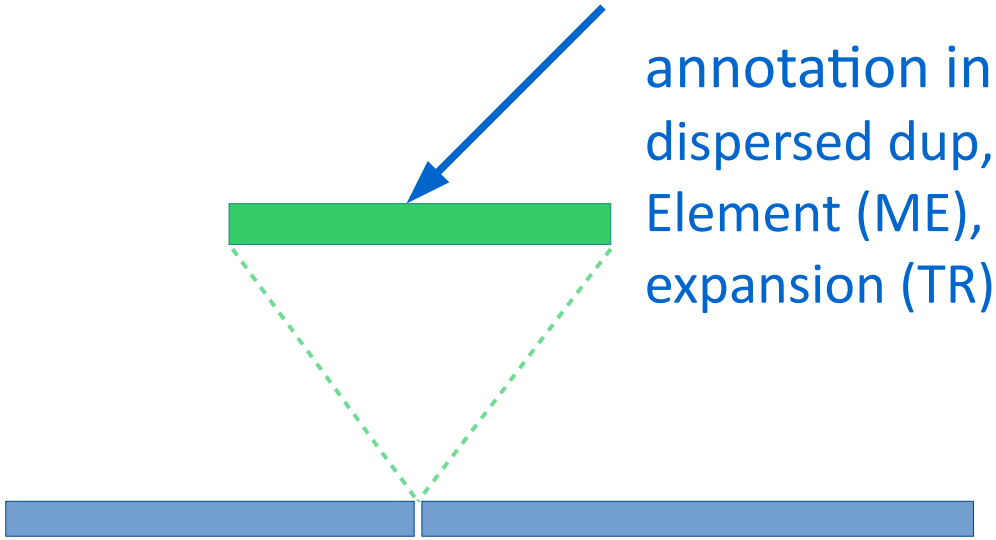
Fine characterization of real human insertions

– Method : 4 levels of characterization

1. size
2. **nature**

Nature of the inserted sequence :

annotation in 5 types (novel, dispersed dup, tandem dup, Mobile Element (ME), Tandem repeat expansion (TR))

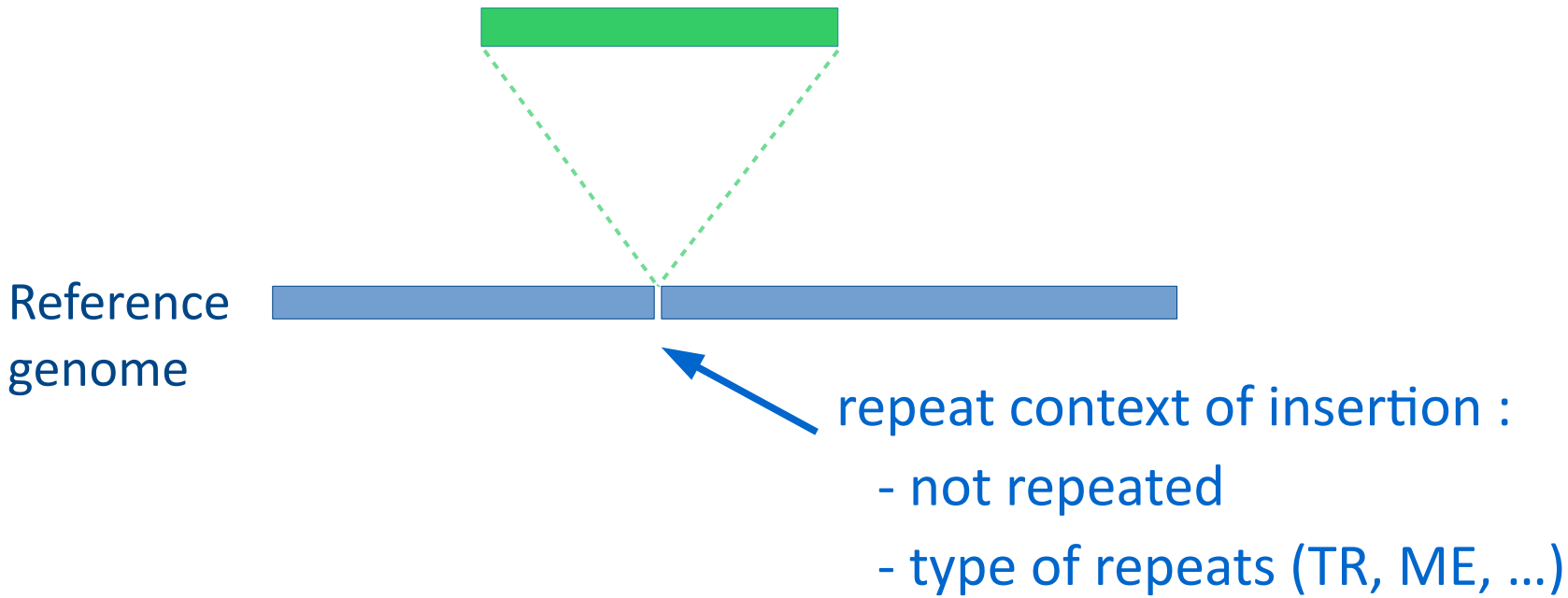


Reference genome

Fine characterization of real human insertions

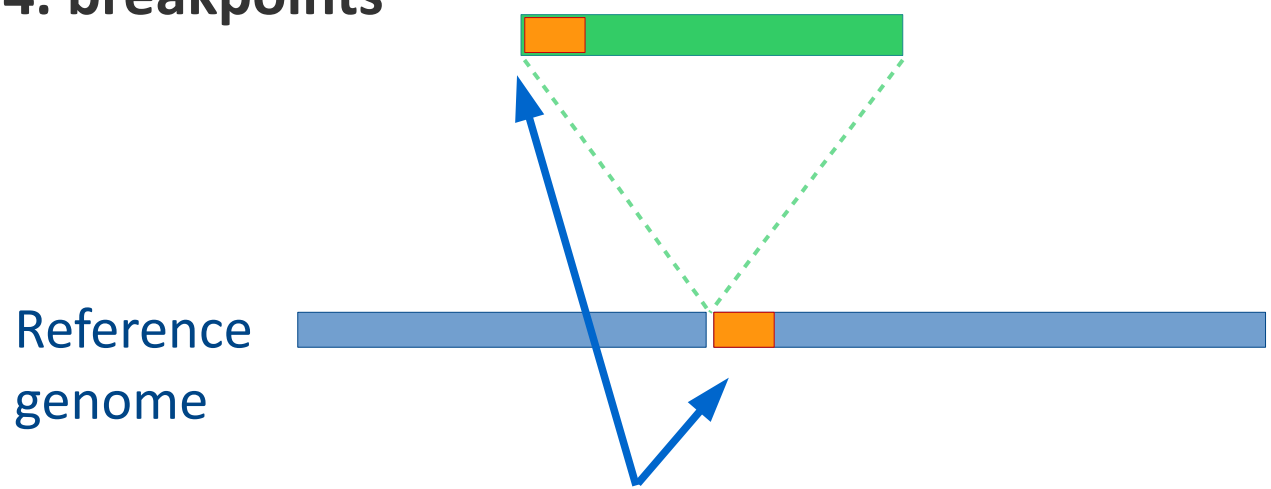
– Method : 4 levels of characterization

1. size
2. nature
3. **genomic location**



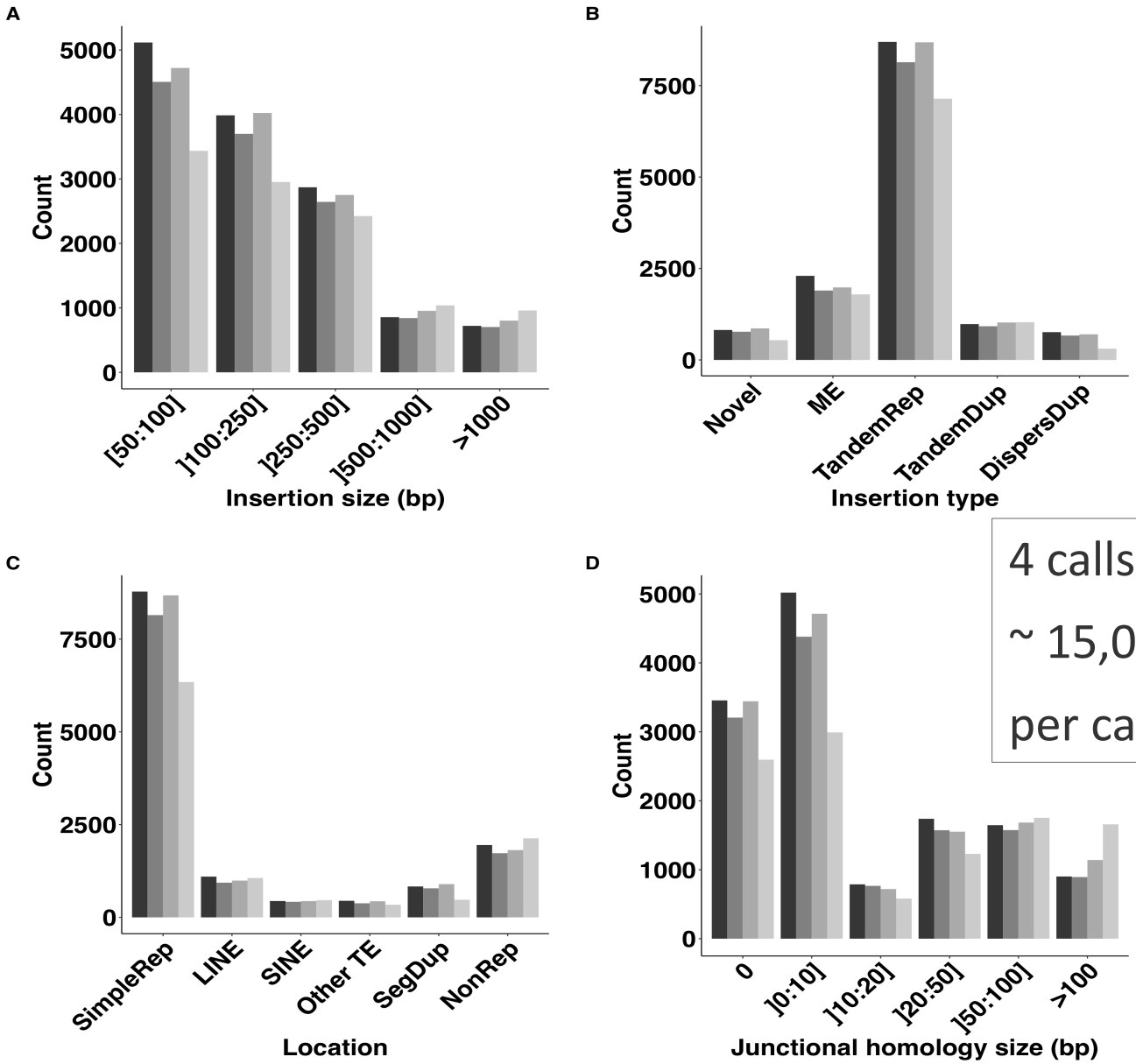
Fine characterization of real human insertions

- Method : 4 levels of characterization
  1. size
  2. nature
  3. genomic location
  4. **breakpoints**



Complexity at breakpoints = size of *junctional homology* (repeat)

Fine characterization of real human insertions

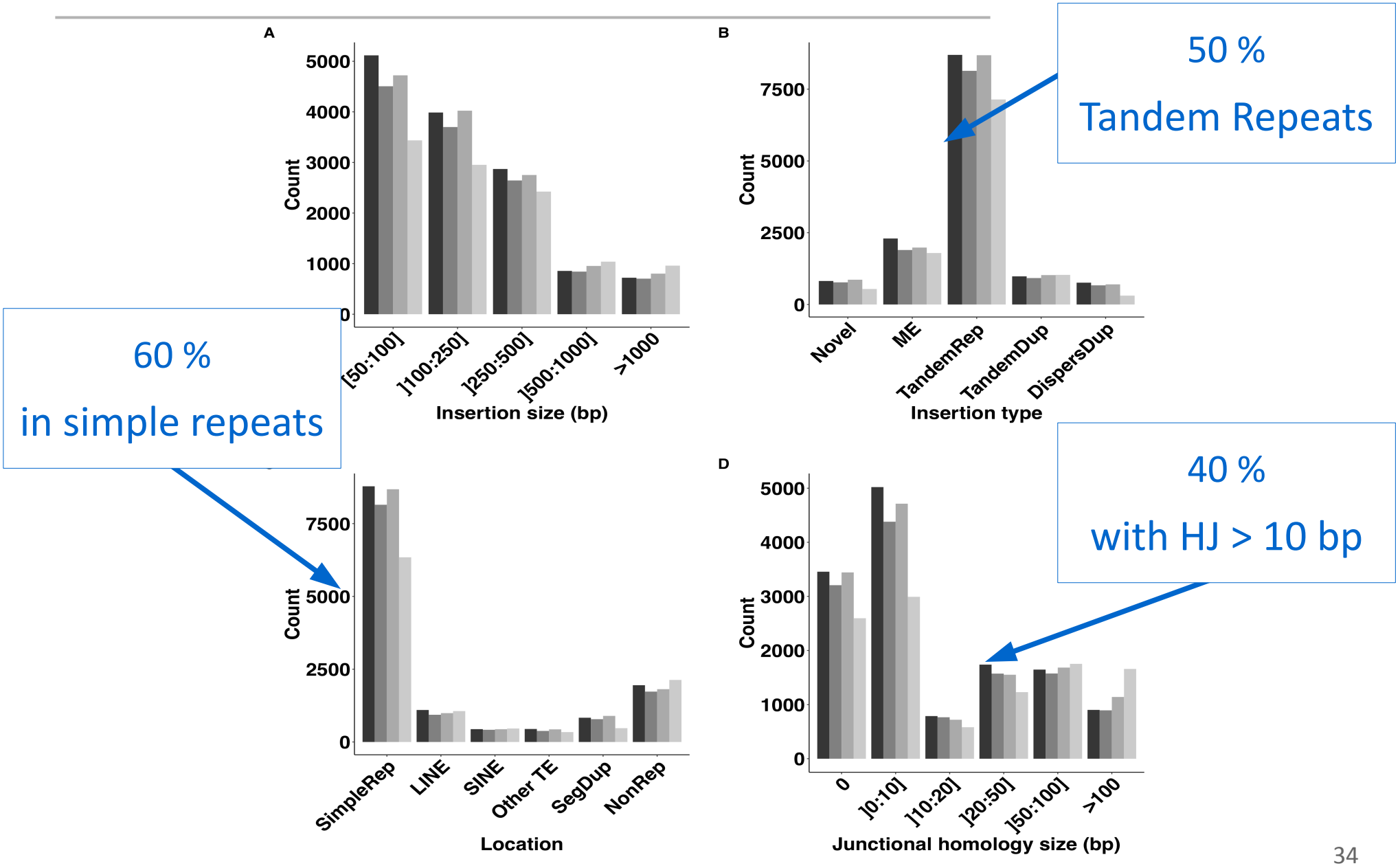


4 callsets with

~ 15,000 insertions

per callset

Most insertions are « difficult »



Which features impact most the recall ?

- Using simulations to disantangle correlated features
  - 1 baseline (easiest case) simulation :
    - 200 simulated insertions on human chr 3
    - size = 250 bp, type = novel, location = exon, HJ = 0 bp
    - 2x150 bp reads at 40x
  - 20 simulated datasets : changing one insertion feature at a time
- Benchmark of 4 insertion callers : [Chen *et al*, 2016; Wala *et al*, 2018 ; Cameron *et al*, 2017]
  - 3 generic latest SV callers (Manta, Svaba, GRIDSS) + MindTheGap
  - Recall : how many of the 200 insertions are discovered and sequence-resolved ?

Benchmark results

	Recall (%)			
	GRIDSS	Manta	SvABA	MTG
Baseline simulation: 250 bp novel sequences in exons	81	100	96	100

Benchmark results

		Recall (%)			
		GRIDSS	Manta	SvABA	MTG
Baseline simulation: 250 bp novel sequences in exons		81	100	96	100
Scenario 1	50 bp	56	100	100	100
Insertion	500 bp	0	0	0	99
size	1,000 bp	0	0	0	98

Benchmark results

		Recall (%)			
		GRIDSS	Manta	SvABA	MTG
Baseline simulation: 250 bp novel sequences in exons		81	100	96	100
Scenario 1 Insertion size	50 bp	56	100	100	100
	500 bp	0	0	0	99
	1,000 bp	0	0	0	98
Scenario 2 Insertion type	Dispersed duplication	0	0	16	96
	Tandem duplication	0	0	0	0
	Mobile element	0	0	61	58
	Tandem repeat (6 bp pattern)	0	0	1	0
	Tandem repeat (25 bp pattern)	0	0	0	0
Scenario 3 Genomic location	No repeat	77	97	93	83
	Simple repeat (<300 bp)	77	98	97	73
	Simple repeat (>300 bp)	77	93	90	58
	SINE	77	99	94	53
	LINE	76	97	95	89
Scenario 4 Junctional homology	10 bp	99	100	92	0
	20 bp	100	100	78	0
	50 bp	6	46	10	0
	100 bp	0	11	0	0
	150 bp	0	0	0	0

Benchmark results

	Recall (%)			
	GRIDSS	Manta	SvABA	MTG
Baseline simulation: 250 bp novel sequences in exons	81	100	96	100
Scenario 1 Insertion size	0 – 100 %			
Scenario 2 Insertion type	0 – 100 %			
Scenario 3 Genomic location	53 – 100 %			
Scenario 4 Junctional homology	0 – 100 %			

High variability in recall between simulations and between tools

Better understanding the loss of recall

- Many different types of insertions
  - A large majority are « difficult »
  - Most impactful features : size and insertion type (Ex : TR)
  - Towards better practices for simulations and benchmarks
- Still some positive findings for improving SV callers :
  - Insertion site often findable, but lack of sequence resolution
  - High variability between tools
    → finding the good combination of SV callers to improve the recall

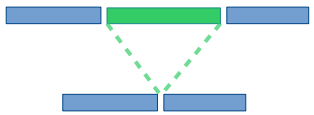
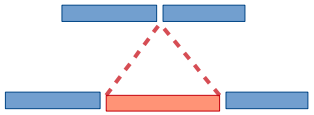
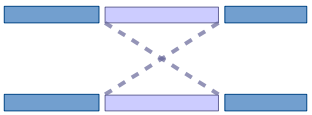
Delage *et al, BMC Genomics*, 2020

In this talk

1. How do we discover SVs with sequencing data ?
2. The case of long insertions
3. After the discovery...

SV genotyping

- Comparing SVs between individuals

			
	0/0	1/1	1/1
	0/1		0/0
	0/0	1/1	0/1

Present/absent ?
Allele quantification

- Importance to distinguish it from the discovery step :
  - SV discovery is prone to FPs and FNs
  - Assessing the absence of a given SV
  - A common SV representation for all compared individuals

State of the art

- History :
  - Before 2018 : no dedicated tools, genotyping is integrated in discovery tools

    → lack of versatility, limited to some SV types
  - Then, multiplication of dedicated methods, ex : SVTyper, SV2, GraphTyper2, Paragraph... (for short reads)
- Methods :
  - Easier than discovery : analyzing mapping signals at pre-defined positions
  - But : mapping to reference only → bias toward the reference allele

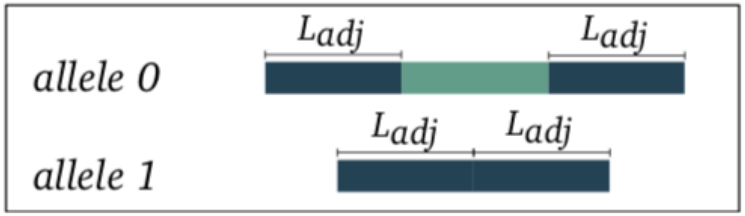
Avoiding the reference bias

- SVJedi : Lecompte *et al*, *Bioinformatics*, 2020

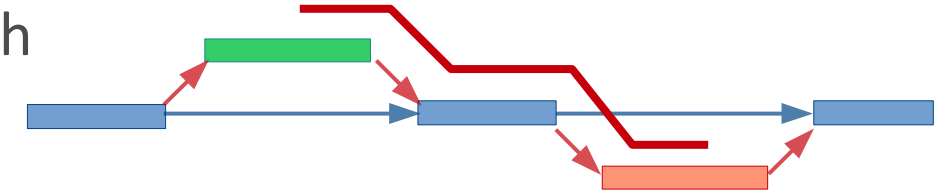
https://github.com/llecompte/SVJedi

BIOCONDA

  - Mapping to both alleles
  - For long reads



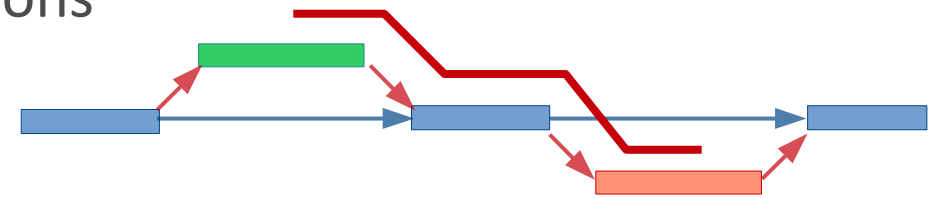
- Graph-based representation
  - One allele/haplotype = one path
  - Allows the representation of complex SVs (with multiple breakpoints), close, imbricated SVs...



  - Ex : VG-toolkit (Garrison 2018), Paragraph (Chen 2019), GraphTyper2 (Eggerston 2019), SVJedi-graph https://github.com/SandraLouise/SVJedi-graph

Conclusion

- Various methodological issues behind SV analysis
  - Depending on SV type and sequencing data
  - Other problems after the discovery : SV representation, comparison, genotyping…
- Current/future challenges :
  - Importance of precise reconstruction of alternative alleles (e.g. through local assembly)
  - Pangenomic graph representations for large population data



Acknowledgments

- Genscale team in Rennes

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

- Work of PhD students

Wesley Delage

Lolita Lecompte

Sandra Romain