



HAL
open science

SVJedi: Structural variation genotyping using long reads

Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier, Claire Lemaitre

► To cite this version:

Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier, Claire Lemaitre. SVJedi: Structural variation genotyping using long reads. HiTSeq 2019 - Conference on High Throughput Sequencing, Jul 2019, Basel, Switzerland. hal-02290884

HAL Id: hal-02290884

<https://inria.hal.science/hal-02290884>

Submitted on 18 Sep 2019

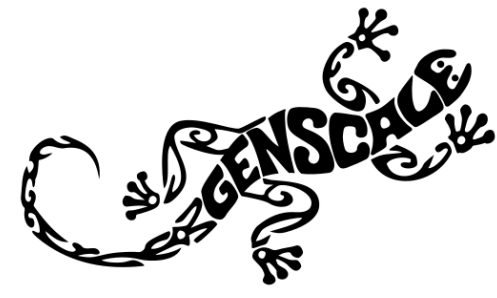
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SVJedi : Structural variation genotyping using long reads

Lolita Lecompte¹, Pierre Peterlongo¹, Dominique Lavenier¹, Claire Lemaitre¹

¹Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France



Introduction

Structural variations (SVs) are DNA segments that have been moved regarding another reference genome. The number of known SVs is increasing, especially with the development of **third generation sequencing technologies**, which produce long reads data.

Discovery vs. Genotyping

SV discovery

Given a reference genome, **discover** SVs between a sequencing sample and a reference genome.

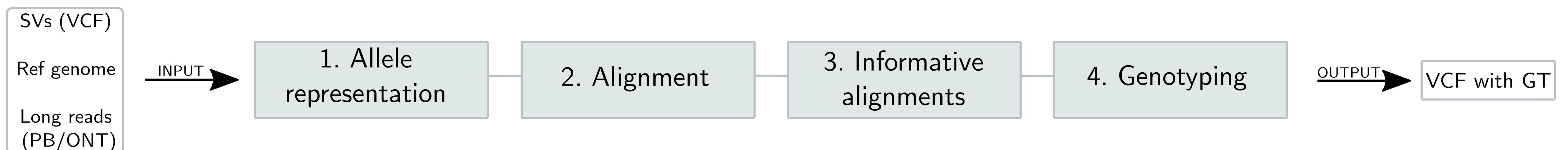
SVIM (2019), Sniffles (2018), pbsv (2018), NanoSV (2017)

SV genotyping

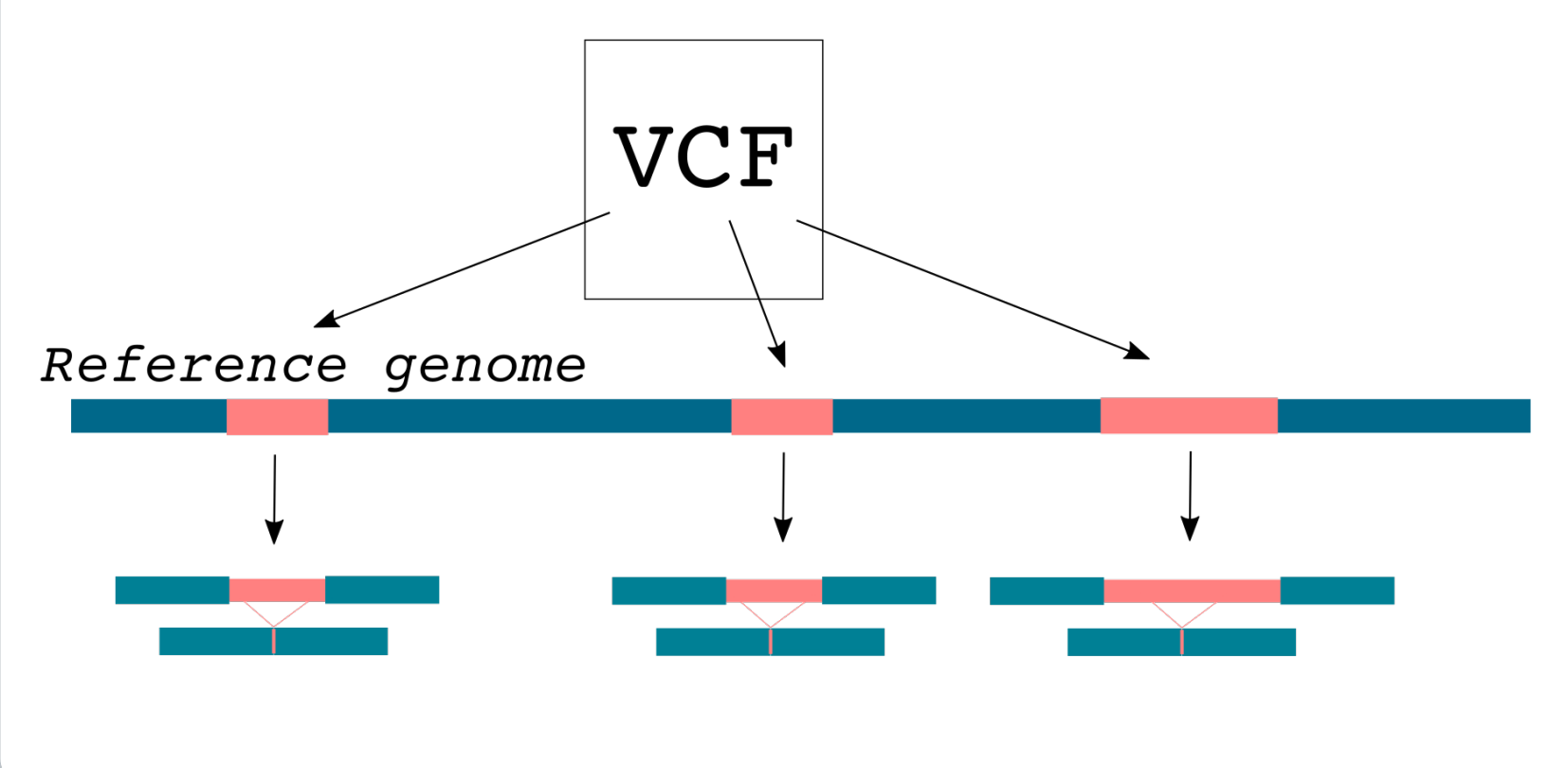
Given a set of already known variants for a species, **assess the presence or absence** of each variant in a given sequencing sample.

There is **no known method** for SV genotyping with long reads.

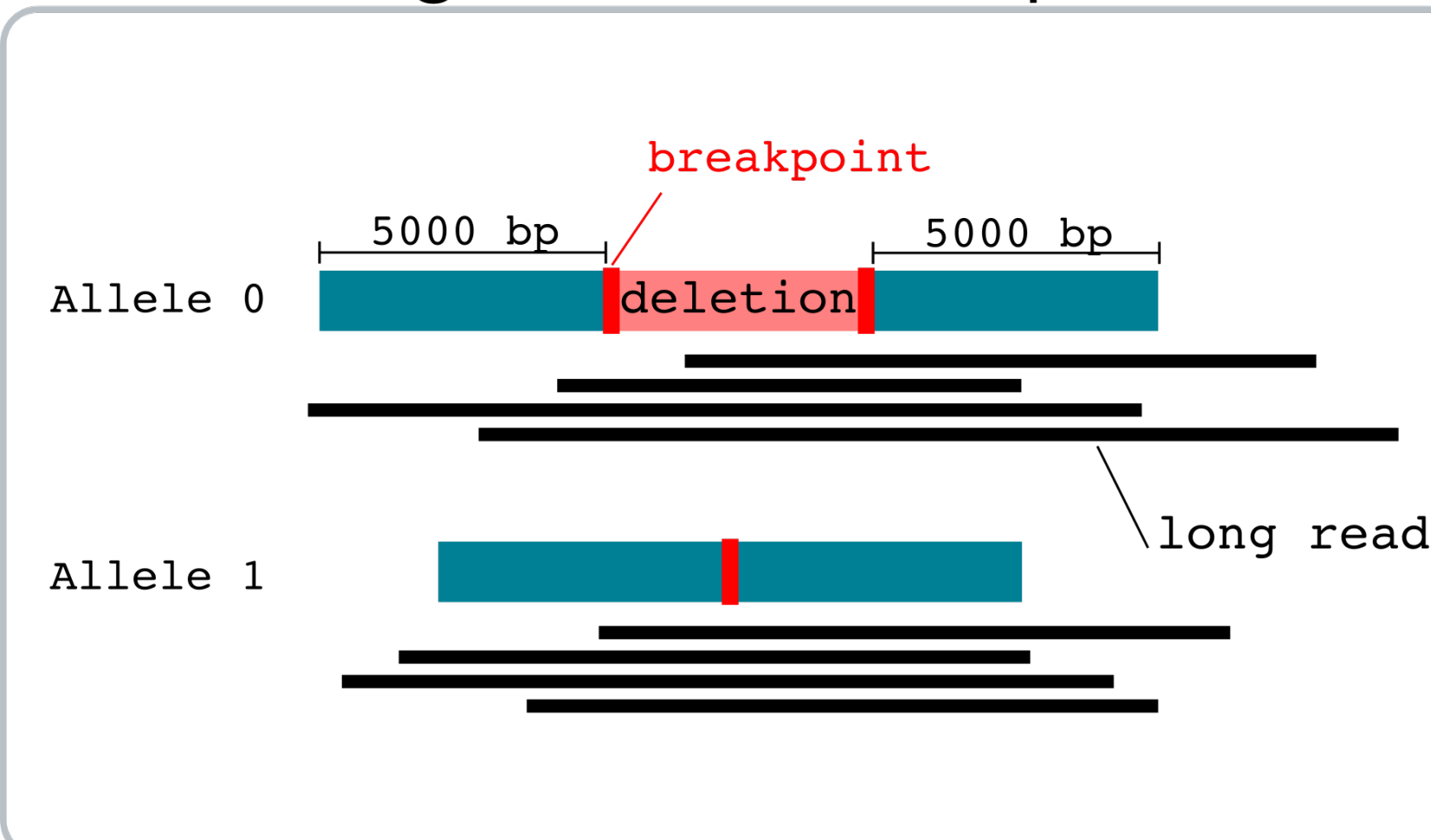
Method - SVJedi



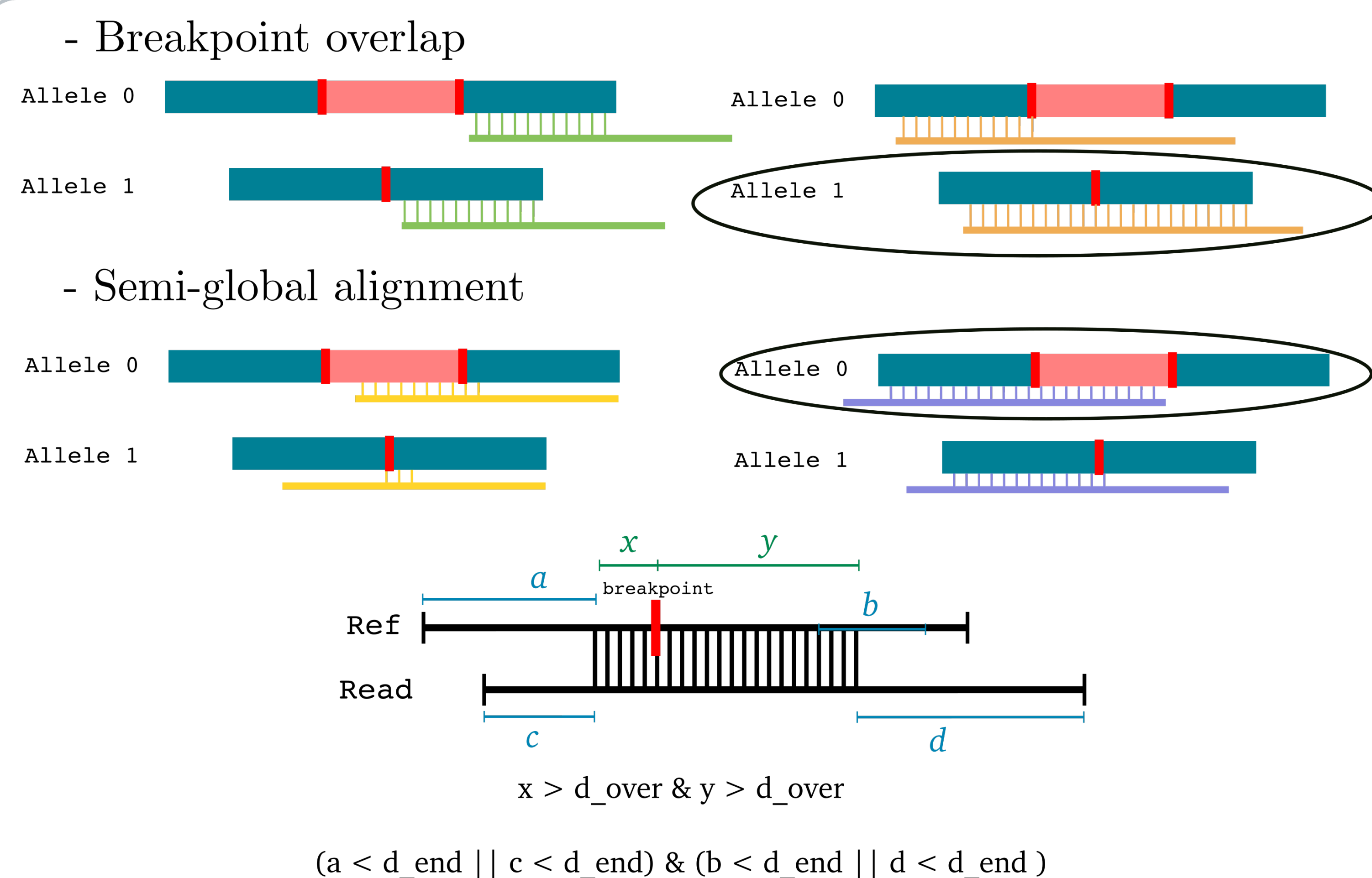
1. Generating allele sequences



2. Alignment : Minimap2



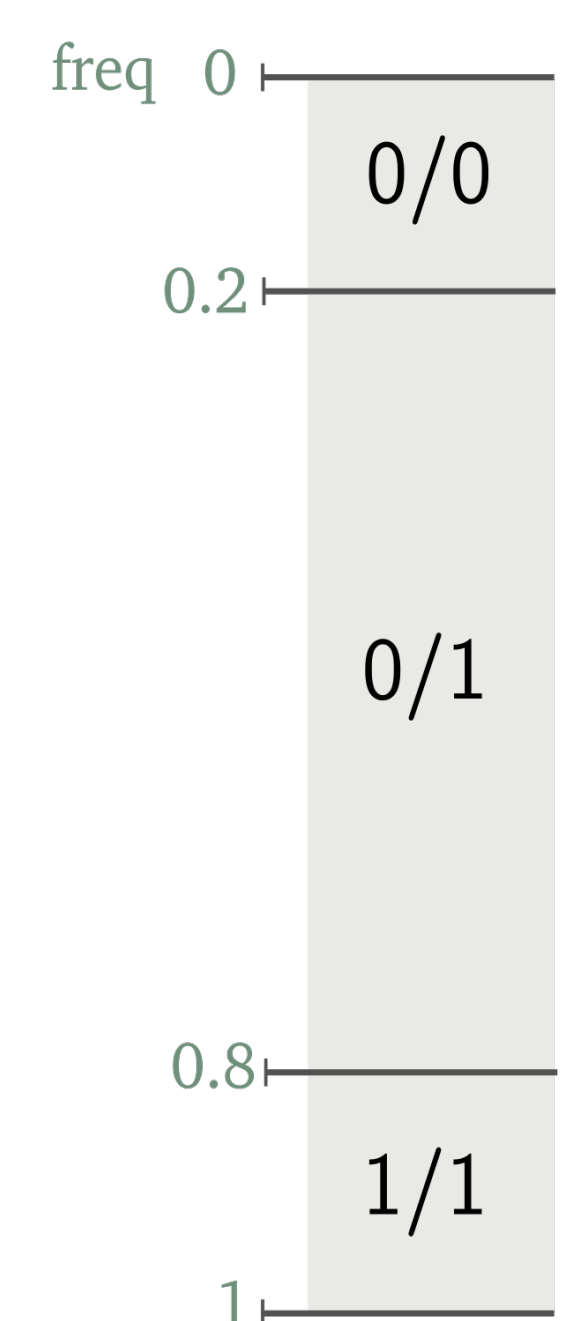
3. Selection of informative alignments



4. Genotype estimation

$$\text{freq} = \frac{\text{allele 0}}{\text{allele 0} + \text{allele 1}}$$

- Normalization
- Minimum read number filter
- Genotype estimation



Results on a simulated dataset

Simulated dataset

- **1,000 deletions** selected on the human chromosome 1 from dbVar (50 bp - 10k bp)
- Equally distributed : 0/0, 0/1, 1/1
- Long reads simulated with SimLoRD [1] 16 % error rate, 20X sequence depth

		SVJedi			
		0/0	0/1	1/1	./.
Truth	0/0	330	3	0	0
	0/1	19	305	6	4
	1/1	3	10	311	9
	./.				

Precision : 95.85 %

Does SV discovery tools correctly estimate genotypes?

		Sniffles [2]		
		0/0	0/1	1/1
Truth	0/1	3	89	0
	1/1	0	254	60
	./.			

Recall : 60.87 %
Precision : 36.70 %

80.9 % of 1/1 deletions are wrongly genotyped as 0/1

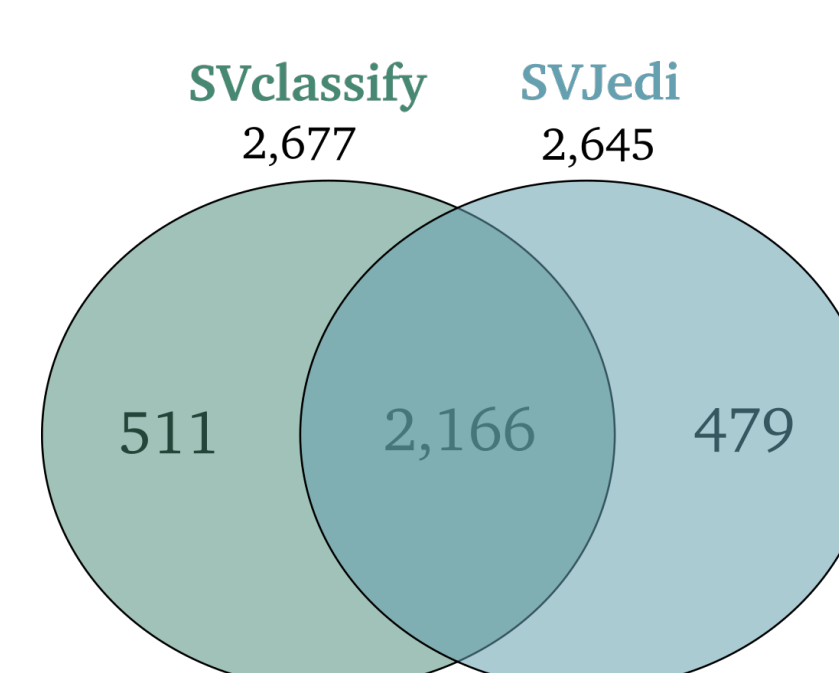
Results on a real dataset

Dataset

- Sample: NA12878
- Data: Nanopore WGS consortium (37X)
- svclassify [3]: **2,677** high confidence deletions

		SVJedi			
		0/0	0/1	1/1	./.
svclassify	0/1	194	1,289	202	28
	1/1	8	75	877	4
	./.				

Identity : 81.89 %
Predicted genotypes = 98.81 %
Runtime : 2h35 (40 CPU)



Conclusion

- ▶ New method to genotype SVs using long reads data.
- ▶ Importance of dedicated genotyping methods.
- ▶ Currently only applied to deletions.
- ▶ Implementation: SVJedi

References

- [1] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. Simlord: simulation of long read data. *Bioinformatics*, 32(17):2704–2706, 2016.
- [2] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, 15(6):461, 2018.
- [3] Hemang Parikh, Marghoob Mohiyuddin, Hugo YK Lam, Hariharan Iyer, Desu Chen, Mark Pratt, Gabor Bartha, Noah Spies, Wolfgang Losert, Justin M Zook, et al. svclassify: a method to establish benchmark structural variant calls. *BMC genomics*, 17(1):64, 2016.

<https://github.com/llecompte/SVJedi>
Contact: lolita.lecompte@inria.fr

