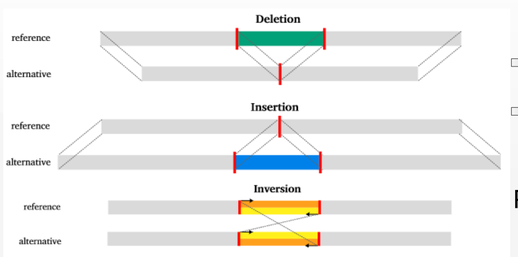
SVJedi-graph: genotyping close and overlapping structural variants with a variation graph and long reads

Sandra Romain[1], Claire Lemaitre[1]

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

Context

Structural variants (SVs) = genomic **rearrangements > 50 bp**.



→ Genome **diversity**

→ Phenotypic variations

↓

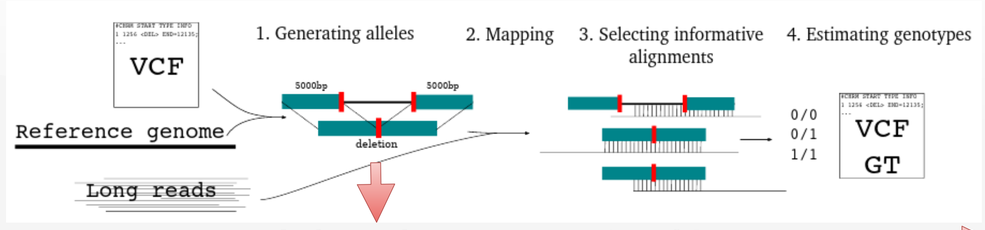
Role in diseases and speciation

SV genotyping = ***"Which alleles are present in sequenced individuals ?"***

Why using a variation graph?

- **Long-read genotyper SVJedi** [1]:

Git: https://github.com/llecompte/SVJedi



independent representation of SVs

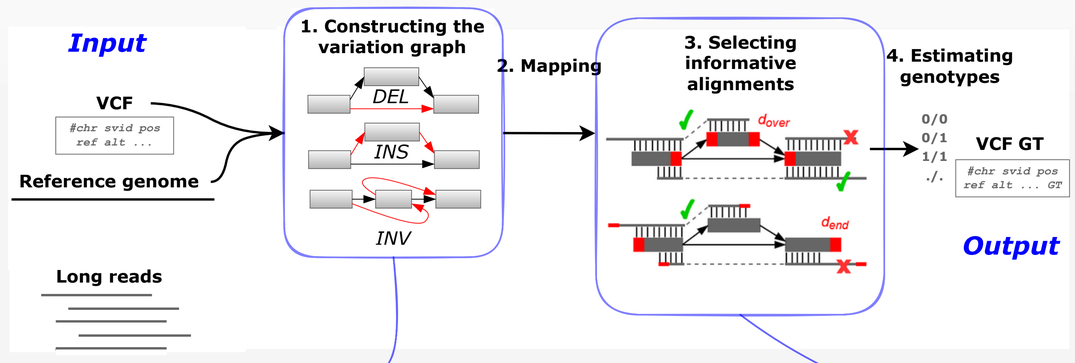
Originality: represents alleles by **partial** genomic sequences.

✓ Better **accuracy**
Lower **computing time**

✗ Hindered genotype estimation for **close SVs**

Method overview

- SVs in the genome represented by a **variation graph**
- Reads mapped with GraphAligner [2]
- Genotype estimation based on **allele coverage** and likelihood estimation



Why long reads?

- SV size: up to **several Mb**.
- SV localization: often in **repeated regions**.

[Key-step] Interpeting & Filtering alignments

- *Aim:* keep **informative** and **non-ambiguous** alignments only

1. Interpreting

Mapping location = **graph node ID**

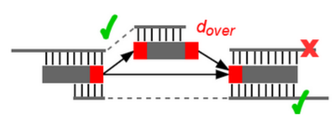
unique number depending on node order construction

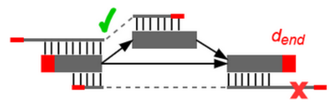
→ **SV(s)** ?

→ **Allele** ?

2. Filtering

- on **breakpoint overlap**
  - *informative on allele ?*





- on **semi-globality**
  - *legitimate ?*

Constructing the variation graph

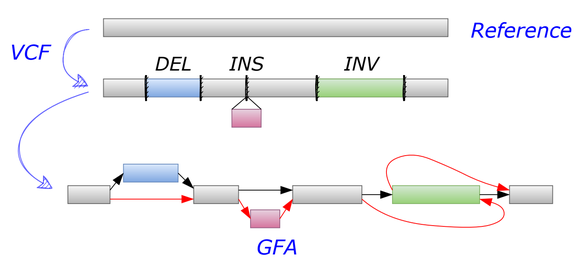
1. Listing SVs breakpoints

Breakpoint

2. Constructing the variation graph

Path of reference allele

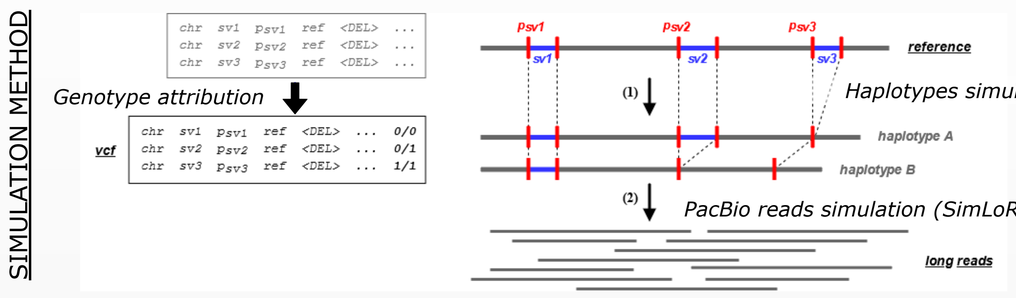
Path of alternative allele



Validation on simulated datasets

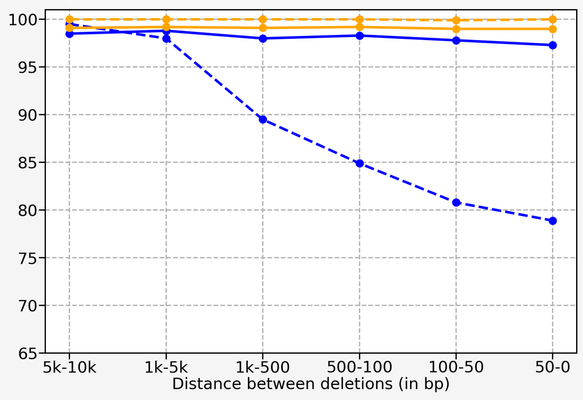
Reference = human chromosome 1 assembly (GRCh37.p13)

SIMULATION METHOD

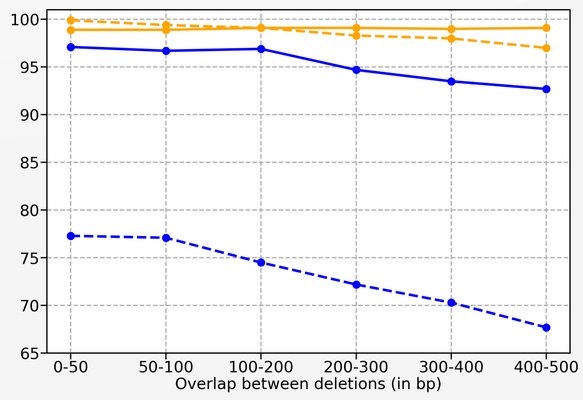


- 6 **close deletions** datasets (decreasing distance ranges)
- 6 **overlapping deletions** datasets (increasing overlapping ranges)

1. Close deletions



2. Overlapping deletions



SVJedi — Genotyping rate

SVJedi-graph — Genotyping accuracy

- High and stable genotyping **accuracy** and **rate** for both close and overlapping SVs

% of accurately predicted genotypes over all predicted genotypes

% of genotyped SVs over all input SVs

Validation on real human datasets (HG002)

1. "Gold standard" SV set

GIAB dataset [4]**:** *(cleaned)*

- human genome reference (GRCh37.p13)
- PacBio reads (HG002)

12,721 SVs → 581 (4.6%) closely located SVs

2. Raw SV callset

Sniffles SV calling → 17,624 SVs → 2,205 (12.5%) closely located SVs

Tool	Genotyping accuracy	Genotyping rate	Time
SVJedi-graph	**92.9**	97.4	15h28m
SVJedi	**92.2**	90.2	2h25m
Sniffles [6] (-Ivcf)	82.0	**99.8**	17h16m
svviz2 [7]	65.9	**100.0**	5days

from [1]

Tool	Genotyping rate
SVJedi-graph	**98.0**
SVJedi	51.0

Conclusions

→ Better accuracy than other long-read SV genotypers.

→ Limitation on SV rich regions solved using a variation graph to represent SVs.

→ Running time still lower than other SVs genotypers (appart from SVJedi).

Available on github (and bioconda soon) (https://github.com/SandraLouise/SVJedi-graph)

References

[1] L. Lecompte, P. Peterlongo, D. Lavenier, and C. Lemaitre. **2020**. SVJedi: genotyping structural variations with long reads. *Bioinformatics*, **36**(17): 4568–4575.
[2] M. Rautiainen and T. Marschall. **2020**. GraphAligner : rapid and versatile sequence-to-graph alignment. *Genome Biology*, **21**(1): 253.
[3] B. K. Stöcker, J. Köster and S. Rahmann. 2016. SimLoRD : Simulation of Long Read Data. Bioinformatics, 32: 2704-2706.
[4] J.M. Zook, N.F. Hansen, N.D. Olson *et al*. **2019**. A robust benchmark for germline structural variant detection. bioRxiv (preprint). doi: https://doi.org/10.1101/664623
[5] T. Jiang, Y. Liu, Y. Jiang *et al*. **2020**. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, **21**(189).
[6] F.J. Sedlazeck, P. Rescheneder, M. Smolka *et al*. **2018**. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, **15**: 461–468.
[7] N. Spies, J.M. Zook, M. Salit *et al*. **2015**. svviz: a read viewer for validating structural variants. *Bioinformatics*, **31**: 3994–3996.

Acknowledgements

This work was supported by the French Agence Nationale de la Recherche [grant number ANR-20-CE02-0017 Divalps].

We are thankful to the Genouest bioinformatics platform for the access to the resources of the Genouest infrastructure.