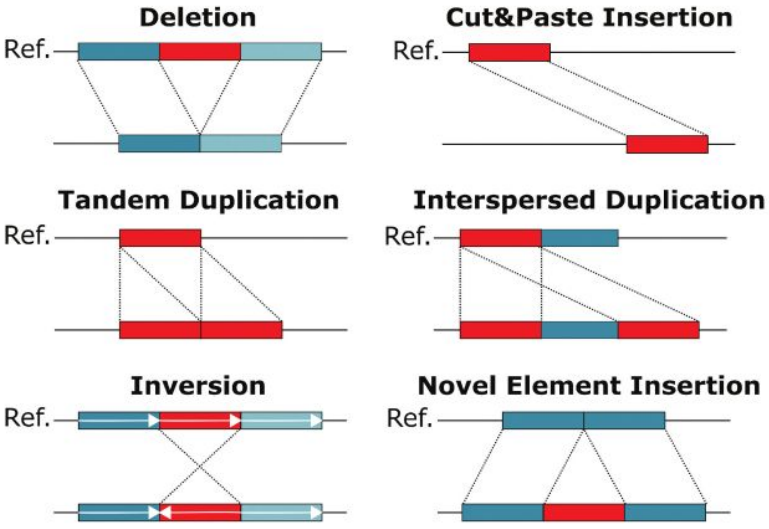
SVJedi-graph: using a variation graph to improve structural variant genotyping with long reads

Sandra ROMAIN [1], Claire Lemaitre [1]

Workshop Data Structures in Bioinformatics - June 2022

INRIA, GenScale team, Rennes, France [1]

Structural variants



Heller and Vingron, 2019

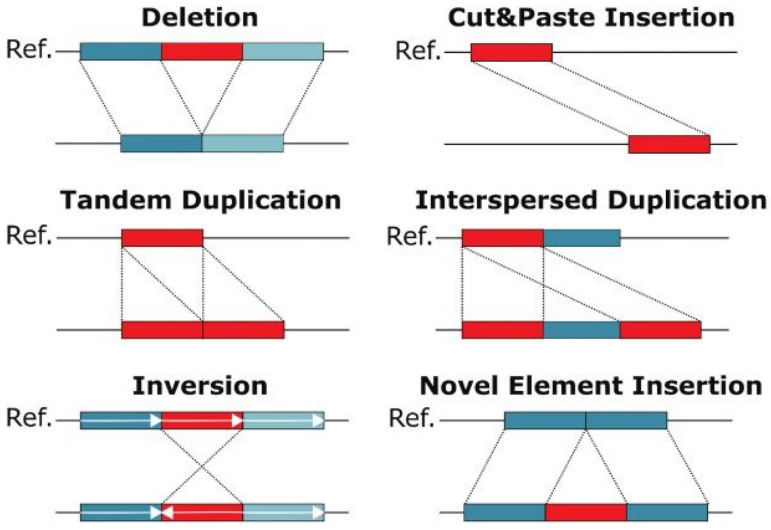
Defined:

as rearrangements ≥ 50 bp

relatively to a reference genome

by
- breakpoints
- sequence

Structural variants



Heller and Vingron, 2019

Defined:

as rearrangements ≥ 50 bp

relatively to a reference genome

by
- breakpoints
- sequence

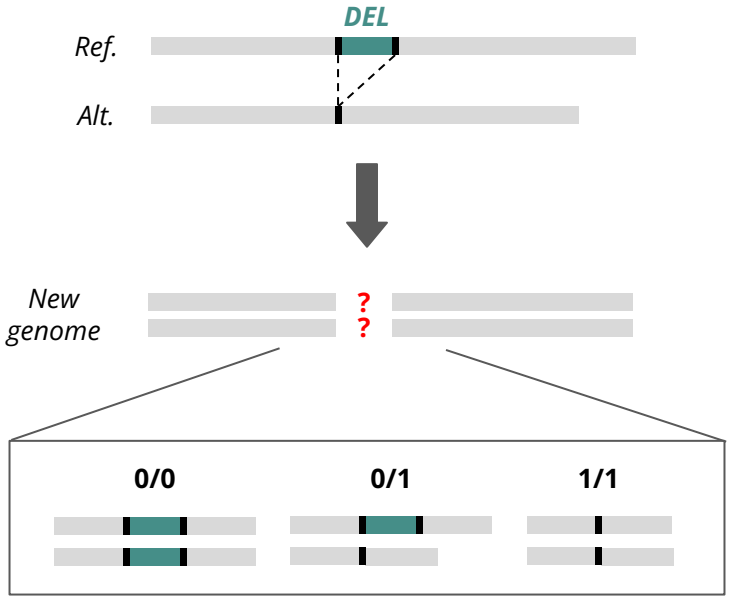
Impact:

depends on genomic context

can lead to
- diseases
- polymorphism in agronomic key traits

Genotyping structural variants

- After SV identification
  - type
  - position
  - sequence for INS
- Presence of the SVs on the haplotypes ?



Approaches:

- Mapping *vs*. “mapping-free”
- Short reads *vs*. long reads

State of the art

Mapping-based genotypers:

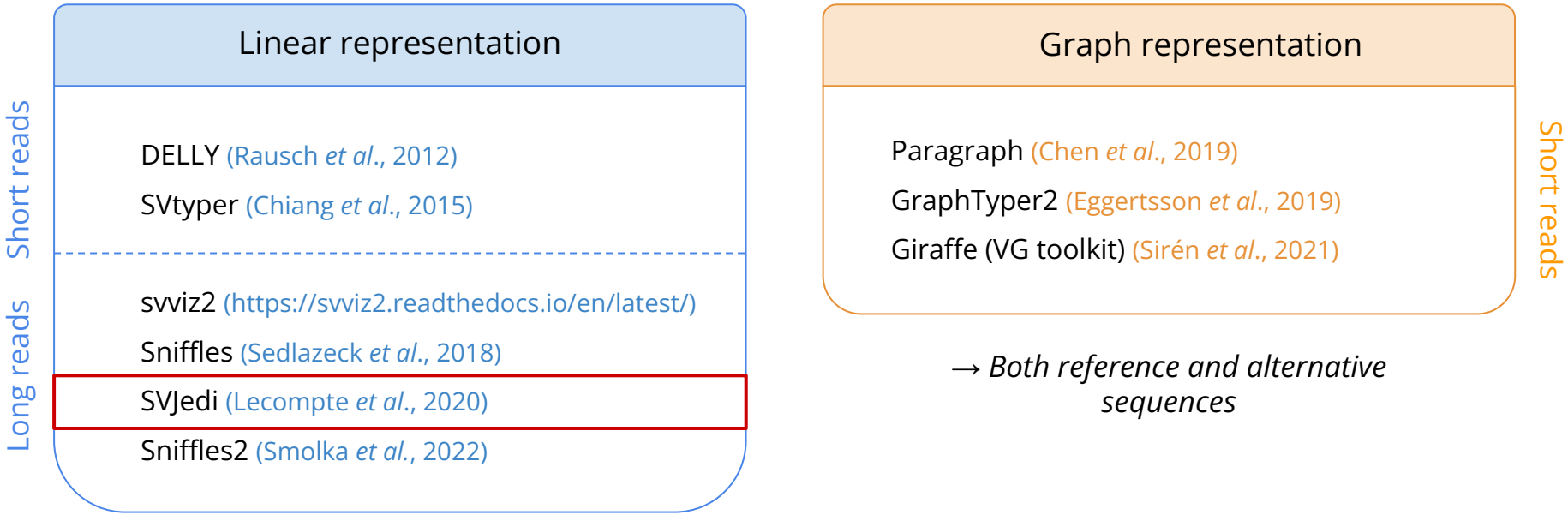
	Linear representation	Graph representation
Short reads	DELLY (Rausch *et al.*, 2012) SVtyper (Chiang *et al.*, 2015)	Paragraph (Chen *et al.*, 2019) GraphTyper2 (Eggertsson *et al.*, 2019) Giraffe (VG toolkit) (Sirén *et al.*, 2021)
Long reads	svviz2 (https://svviz2.readthedocs.io/en/latest/) Sniffles (Sedlazeck *et al.*, 2018) SVJedi (Lecompte *et al.*, 2020) Sniffles2 (Smolka *et al.*, 2022)	

→ *Reference bias*

→ *Both reference and alternative sequences*

State of the art

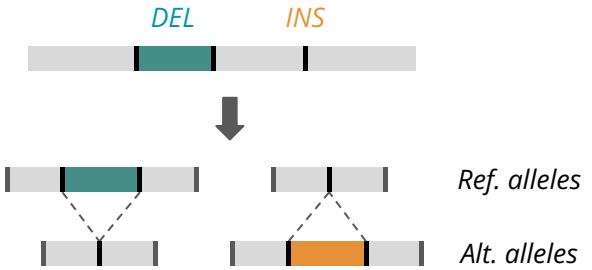
Mapping-based genotypers:



→ *Reference bias*

→ *Both reference and alternative sequences*

SVJedi (Lecompte *et al.*, 2020)



Principle: Representing both alleles for each SV in linear reference

→ *Reduce reference bias*

Tool	Genotyping accuracy	Genotyping rate	Time
SVJedi	**92.2**	90.3	**2h25m**
Sniffles -Ivcf	82.0	**99.8**	17h16m
svviz2	65.9	**100**	5days
Sniffles (discovery mode)	43.6	48.1	18h04m
pbsv	77.9	65.3	5h29m

from Lecompte et al., 2020

Limitation: Drop of genotyping rate with close/overlapping SVs

→ ⚠ *Sequence redundancy*

https://github.com/llecompte/SVJedi

Our contribution: SVJedi-graph

Long read SV genotyper using a variation graph representation

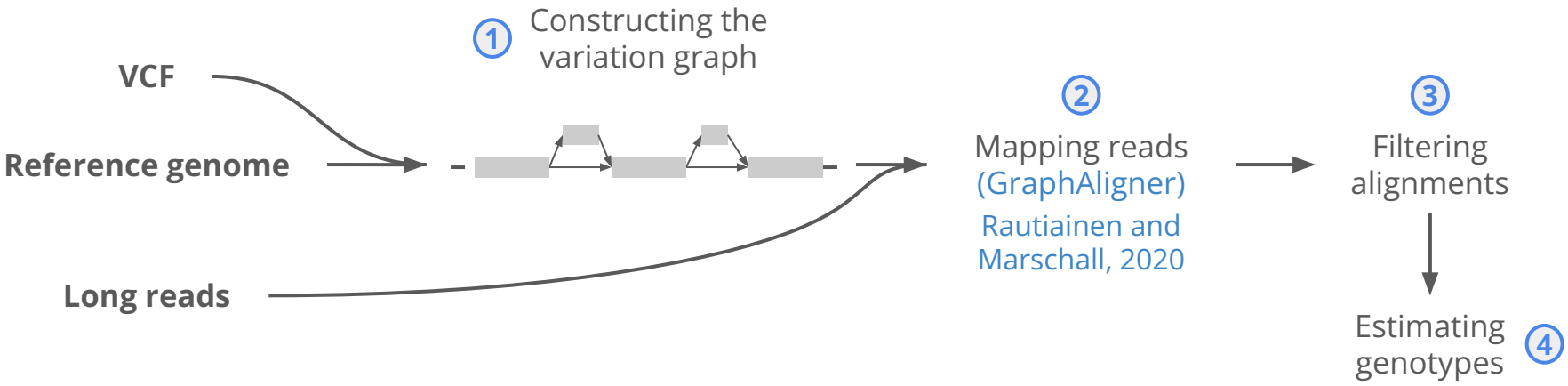
- Improve close SV genotyping by using a variation graph

- Represent the whole genome sequence

Method

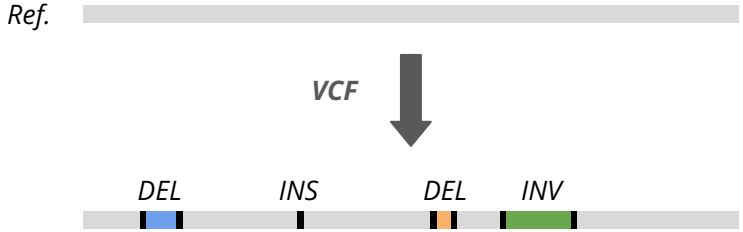
Overview of SVJedi-graph

Input: reference genome, SV set, long reads



Output: genotyped SV set

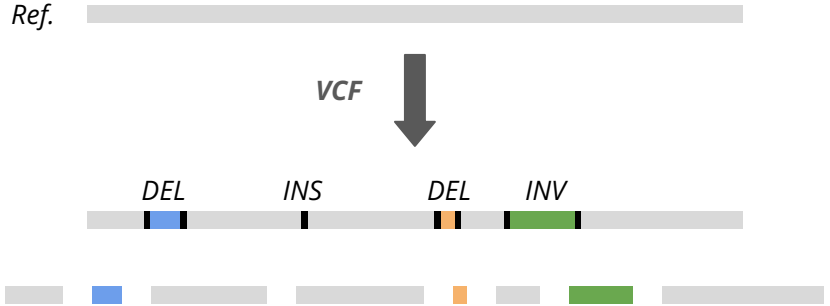
(1) Constructing the variation graph



For each chromosome:

1. List & sort breakpoint positions

(1) Constructing the variation graph

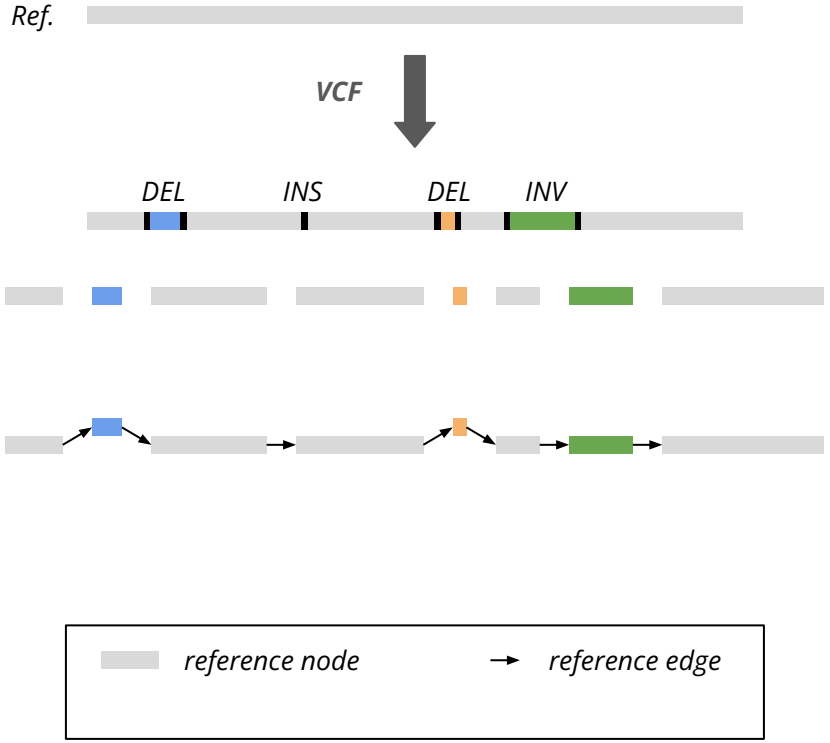


For each chromosome:

1. List & sort breakpoint positions
2. Split sequence at each breakpoint

1 fragment = 1 node

(1) Constructing the variation graph

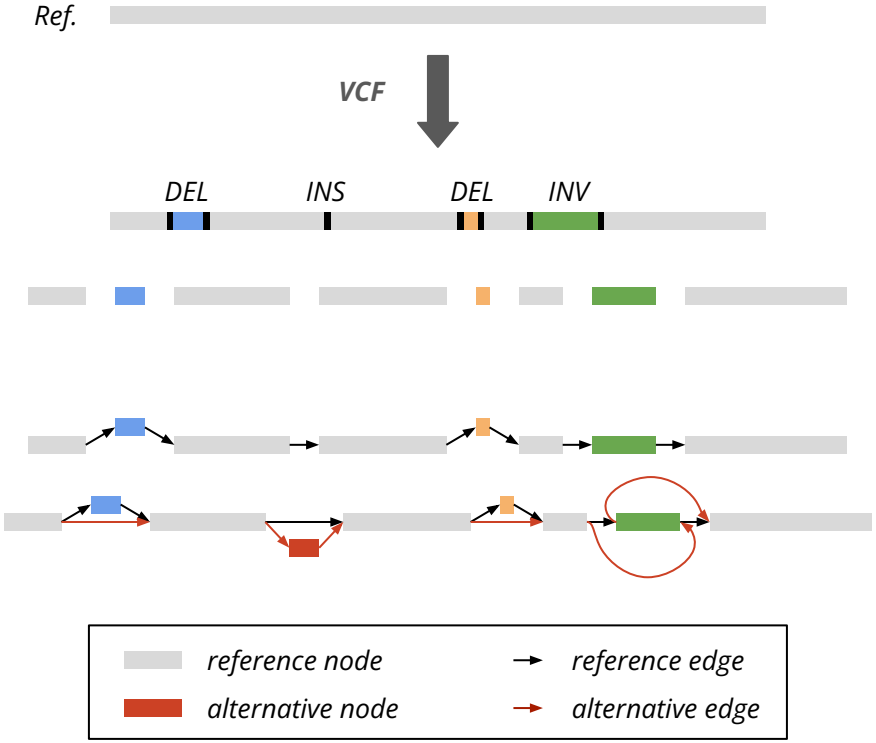


For each chromosome:

1. List & sort breakpoint positions
2. Split sequence at each breakpoint

   1 fragment = 1 node
3. Add reference edges

(1) Constructing the variation graph



For each chromosome:

1. List & sort breakpoint positions
2. Split sequence at each breakpoint

   1 fragment = 1 node
3. Add reference edges
4. Add alternative edges

   + alternative nodes for insertions

(2)(3) Mapping the reads and filtering the alignments

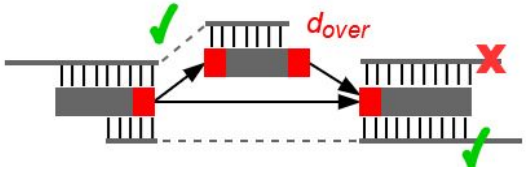
Mapping: GraphAligner (Rautiainen and Marschall, 2020)

Alignments filters:

➢ Number of nodes in the alignment path >= 2 → *Filtering alignments to analyse*

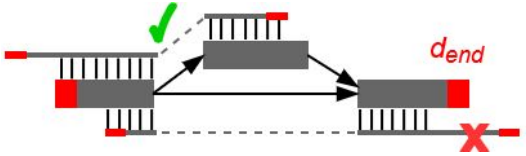
➢ Breakpoints overlap

→ *Confidence in supported allele*



➢ Alignment semi-globality

→ *Accuracy of mapping location*



(4) Predicting the genotype

- Count supporting reads for each allele
- Normalize by allele length ratio
- Compute likelihood for each genotype

$$\ell(0/0) = (1 - err)^{c_0^*} \times err^{c_1} \times C_{c_0^*+c_1}^{c_0^*}$$

$$\ell(1/1) = err^{c_0^*} \times (1 - err)^{c_1} \times C_{c_0^*+c_1}^{c_0^*}$$

$$\ell(0/1) = \left(\tfrac{1}{2}\right)^{c_0^*+c_1} \times C_{c_0^*+c_1}^{c_0^*}$$

Reused from SVJedi (Lecompte et al., 2020)

Evaluation on simulated datasets

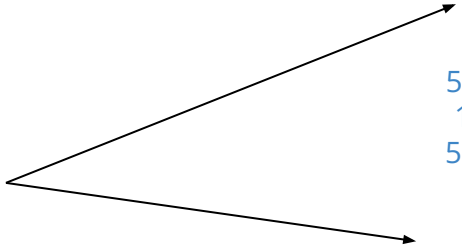
The simulated datasets

Reference: human chromosome 1 (GRCh37.p13)

SV sets generation:

- ➢ 1,000 deletions from dbVar
- ➢ + close/overlapping deletions

  (⅓ 0/0 - ⅓ 0/1 - ⅓ 1/1)



6 sets w/ close deletions

5k → 10k	100 → 500
1k → 5k	50 → 100
500 → 1k	0 → 50

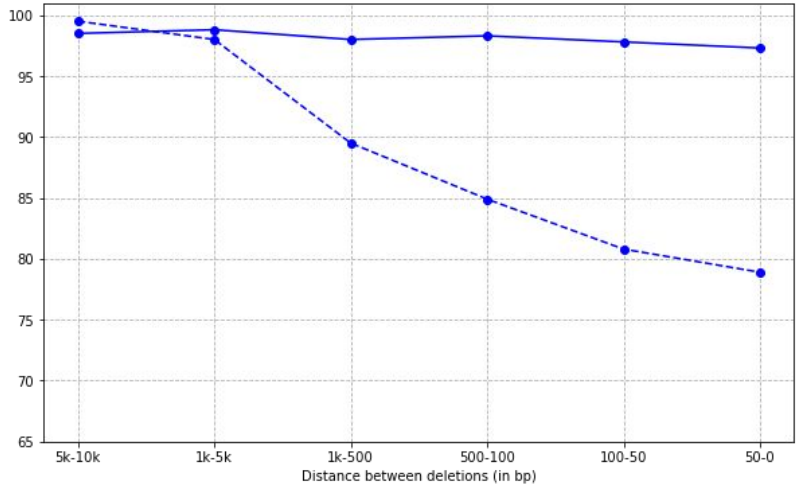
6 sets w/ overlapping deletions

0 → 50	200 → 300
50 → 100	300 → 400
100 → 200	400 → 500

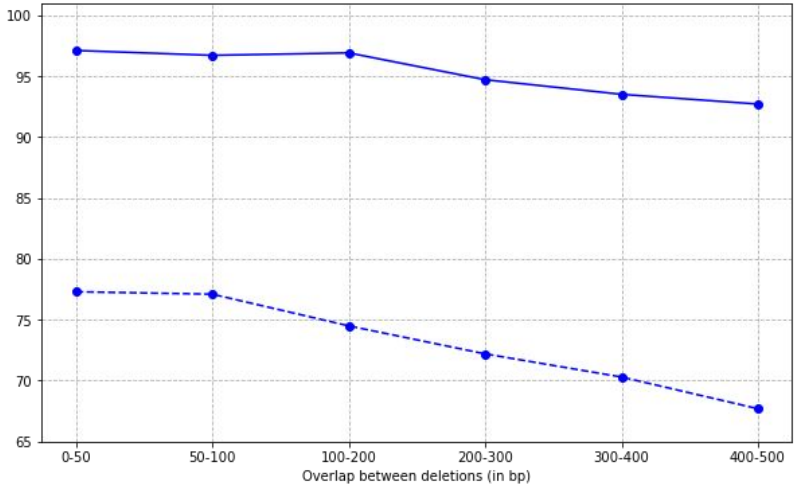
Reads simulation: PacBio, 16 % error rate (SimLoRD)

The simulated datasets - Results (SVJedi)

Close deletions



Overlapping deletions



genotyping accuracy

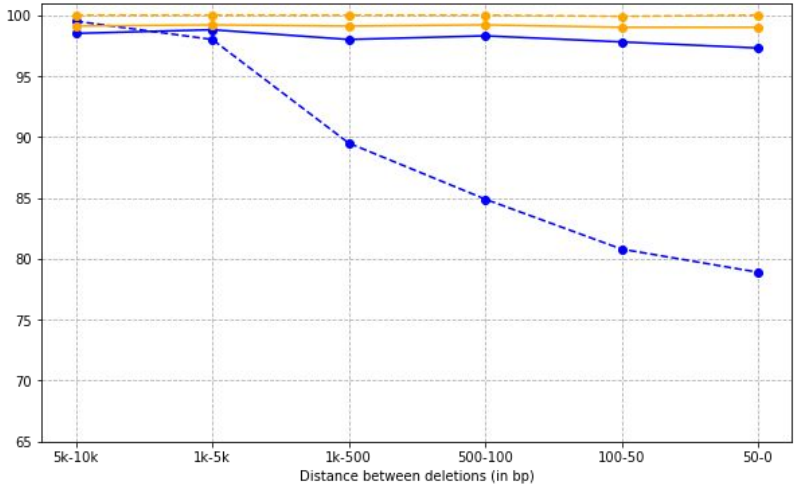
genotyping rate

Rate: % of SVs genotyped / all SVs

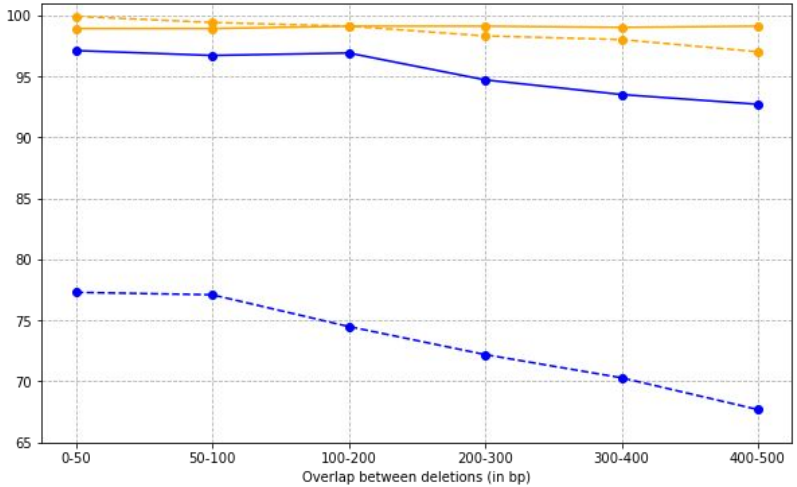
Accuracy: % of SVs accurately genotyped / genotyped SVs

The simulated datasets - Results (SVJedi-graph)

Close deletions



Overlapping deletions



genotyping accuracy

genotyping rate

SVJedi

SVJedi-graph

➢ **Recovery of genotyping rate**

Evaluation on real dataset

The GIAB dataset

Reference: human reference genome (GRCh37.p13)

Reads: PacBio from HG002 (GIAB dataset)

SV set: HG002 Tier 1 (Zook *et al.*, 2019)

- ➢ 5,464 deletions
- ➢ 7,281 insertions

} *with ground truth genotypes*

The GIAB dataset - Results

Reference: human reference genome (GRCh37.p13)

Reads: PacBio from HG002 (GIAB dataset)

SV set: HG002 Tier 1 (Zook *et al.*, 2019)

- ➢ 5,464 deletions
- ➢ 7,281 insertions

} *with ground truth genotypes*

Tool	Genotyping accuracy	Genotyping rate	Time
SVJedi	**92.2**	90.3	**2h25m**
Sniffles -Ivcf	82.0	**99.8**	17h16m
svviz2	65.9	**100**	5days
Sniffles (discovery mode)	43.6	48.1	18h04m
pbsv	77.9	65.3	5h29m
SVJedi-graph	**92.9**	97.4	15h28m

→ *Time cost of mapping on graph*

New SV calling dataset from GIAB data

Reference: human reference genome (GRCh37.p13)

Reads: PacBio from HG002 (GIAB dataset)

SV calling: NGMlr + Sniffles (Sedlazeck *et al*., 2018)

- ➢ 7,922 deletions
- ➢ 9,529 insertions
- ➢ 202 inversions

Dataset	all SVs	“close” SVs
GIAB “gold standard”	12,721	581 (4.6%)
New SV calling set	**17,624**	**2,205 (12.5%)**

New SV calling dataset from GIAB data - Results

Reference: human reference genome (GRCh37.p13)

Reads: PacBio from HG002 (GIAB dataset)

SV calling: NGMlr + Sniffles (Sedlazeck *et al*., 2018)

- ➢ 7,922 deletions
- ➢ 9,529 insertions
- ➢ 202 inversions

Dataset	all SVs	“close” SVs
GIAB “gold standard”	12,721	581 (4.6%)
New SV calling set	**17,624**	**2,205 (12.5%)**

	Genotyping rate
SVJedi	51 %
SVJedi-graph	**98 %**

Concluding remarks

Implemented in python

Availability:

 https://github.com/SandraLouise/SVJedi-graph

 (soon)

Work in progress:

Evaluating genotyping accuracy on the GIAB dataset

Improve read mapping time

Genotyping translocations

Acknowledgements

Mobility grant

Access to computing cluster

And my PhD supervisors: Claire Lemaitre and Fabrice Legeai

anr®

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

References (1)

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R.M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D.R., Schatz, M.C., Sedlazeck, F.J. and Eberle, M.A.. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, **20**(291) (**2019**). https://doi.org/10.1186/s13059-019-1909-7

Chiang, C., Layer, R., Faust, G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R. and Hall, I.R.. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nature Methods*, **12**: 966–968 (**2015**). https://doi.org/10.1038/nmeth.3505

Eggertsson, H.P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M.T., Gudbjartsson, D.F., Stefansson, K., Halldorsson, B.V. and Melsted, P.. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, **10**(5402) (**2019**). https://doi.org/10.1038/s41467-019-13341-9

Heller, D., Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics*, **35**: 2907–2915 (**2019**). https://doi.org/10.1093/bioinformatics/btz041

Lecompte, L., Peterlongo, P., Lavenier, D., Lemaitre, C., SVJedi: genotyping structural variations with long reads. *Bioinformatics*, **36**(17): 4568–4575 (**2020**). https://doi.org/10.1093/bioinformatics/btaa527

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., Korbel, J.O.. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**: 333–i339 (**2012**). https://doi.org/10.1093/bioinformatics/bts378

Rautiainen, M., Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, **21**(253) (**2020**). https://doi.org/10.1186/s13059-020-02157-2

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A. and Schatz, M.C.. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, **15**: 461–468 (**2018**). https://doi.org/10.1038/s41592-018-0001-7

Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T.W., Ratan, A., Taylor, K.D., Rich, S.S., Rotter, J.I., Haussler, D., Garrison, E., Paten, B.. Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *bioRxiv* (**2020**). doi: https://doi.org/10.1101/2020.12.04.412486

References (2)

Smolka, M., Paulin, L.F., Grochowski, C.M., Mahmoud, M., Behera, S., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S.W., Carvalho, C.M.B., Proukakis, C., Sedlazeck, F.J.. Comprehensive Structural Variant Detection: From Mosaic to Population-Level. *bioRxiv* (**2022**). doi: https://doi.org/10.1101/2022.04.04.487055

Spies, N., Zook, J.M., Salit, M., Sidow, A.. svviz: a read viewer for validating structural variants. *Bioinformatics*, **31**(24): 3994–3996 (**2015**). doi: https://doi.org/10.1093/bioinformatics/btv478

Stöcker, B.K., Köster, J., Rahmann, S.. SimLoRD: Simulation of Long Read Data. *Bioinformatics*, **32**(17): 2704–2706 (**2016**). https://doi.org/10.1093/bioinformatics/btw286

Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C., Sahraeian, S.M.E., Huang, V., Rouette, A., Alexander, N., Mason, C.E., Hajirasouliha, I., Ricketts, C., Lee, J., Tearle, R., Fiddes, I.T., Martinez-Barrio, A., Wala, J., Carroll, A., Ghaffari, N., Rodriguez, O.L., Bashir, A., Jackman, S., Farrell, J.J., Wenger, A.M., Alkan, C., Soylev, A., Schatz, M.C., Garg, S., Church, G., Marschall, T., Chen, K., Fan, X., English, A.C., Rosenfeld, J.A., Zhou, W., Mills, R.E., Sage, J.M., Davis, J.R., Kaiser, M.D., Oliver, J.S., Catalano, A.P., Chaisson, M.J.P., Spies, N., Sedlazeck, F.J. and Salit, M.. A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*, **38**: 1347–1355 (**2020**). https://doi.org/10.1038/s41587-020-0538-8