Identification, genotyping and representation of structural variants in pangenomes

PhD defense of ***Sandra ROMAIN***
November 8th, 2024

Reviewers: **Birte KEHR**, Professor, University of Regensburg (Germany)
François SABOT, Research Director, IRD Montpellier

Examiners: **Sèverine BERARD**, Senior Lecturer, Université de Montpellier
Pierre PETERLONGO, Research Director, Inria - Univ. Rennes

Thesis supervisors: **Claire LEMAITRE**, Research Director, Inria - Univ. Rennes
Fabrice LEGEAI, Research Engineer, IGEPP INRAe Rennes

The complex of alpine *Coenonympha* butterfly species

- **Divalps** ANR project
- 4 species
- Hybridization
  - 2 **parental** species
  - 2 **hybrid** species

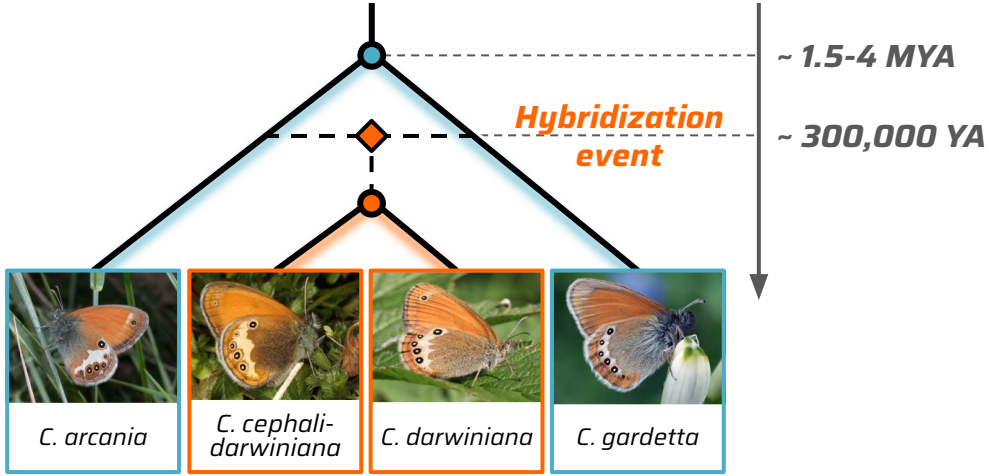
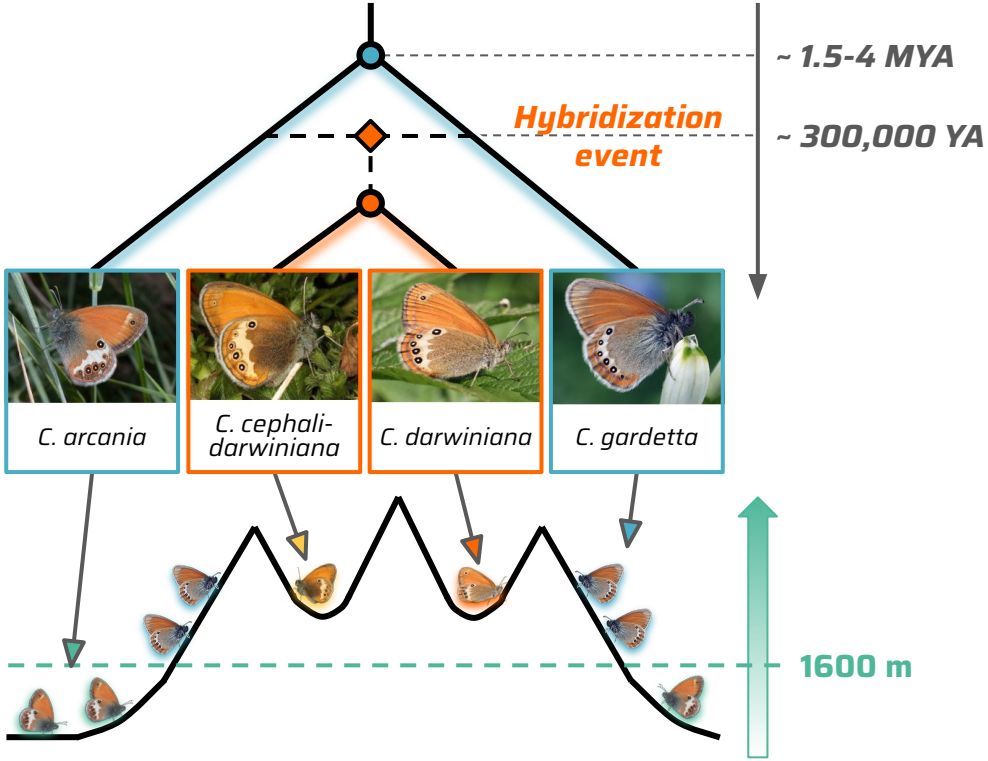


Photo credits: Daniel Morel (*C. arcania*), Wolfgang Wagner (*C. cephalidarwiniana*), Matt Rowlings (*C. darwiniana*), Claire Hoddé (*C. gardetta*)

The complex of alpine *Coenonympha* butterfly species

- **Divalps** ANR project
- 4 species
- Hybridization
  - 2 **parental** species
  - 2 **hybrid** species
- **Altitudinal gradient**
  - climate, O^2
- **Few contact zones** between hybrid and parental species



Genomic factors involved in speciation, adaptation?

Previous genetic diversity analyses

➔ **Previous data:**

- short reads (150 bp) → **punctual** genomic variations (1 bp)
- **gene flow** observed between species
- but **high divergence**

⇨ **Structural differences between the genomes of the species?**

Previous genetic diversity analyses

➔ **Previous data:**

- short reads (150 bp) → **punctual** genomic variations (1 bp)
- **gene flow** observed between species
- but **high divergence**

⇨ **Structural differences between the genomes of the species?**

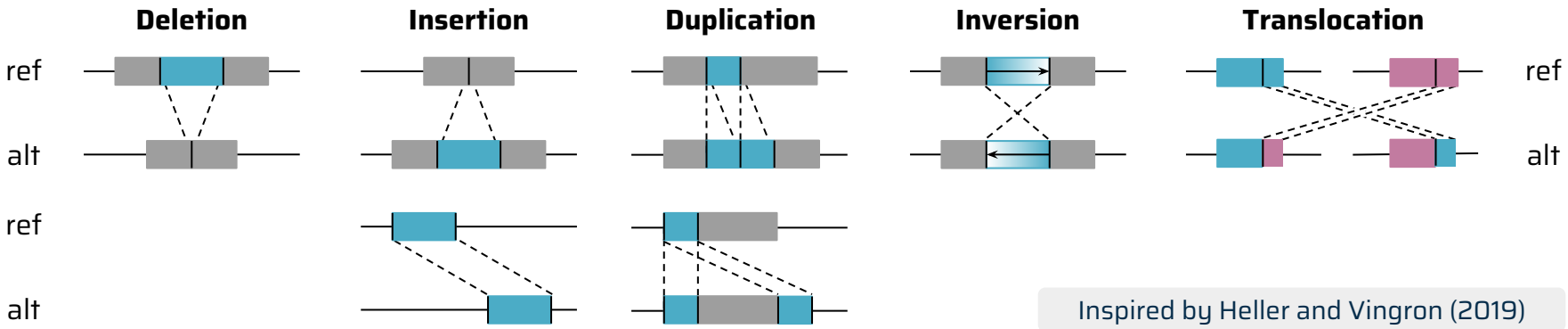
➔ **New data:**

- high-quality long reads (~15 kbp) → access to genome structure

My work: characterization of the structural variations in *Coenonympha* sp. genomes

Structural Variants (SVs)

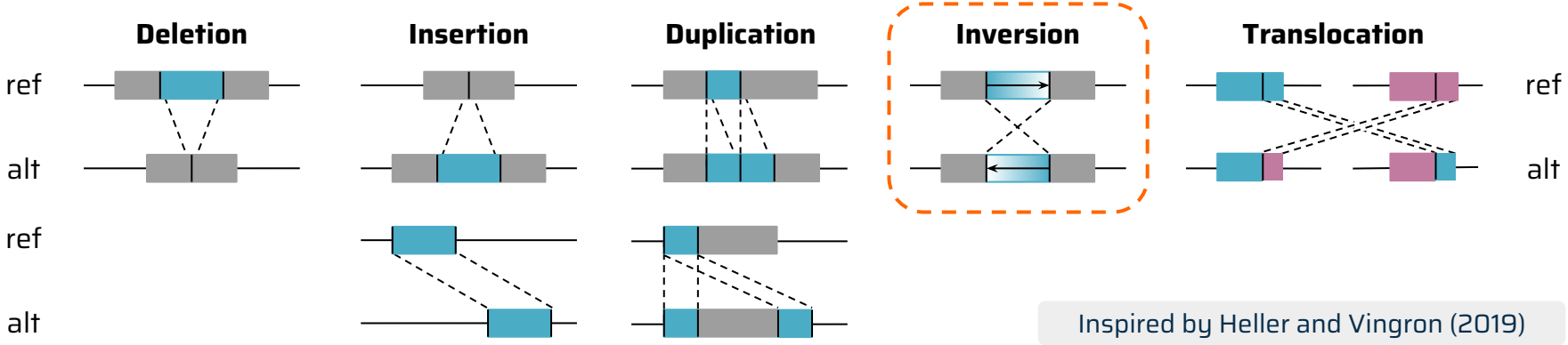
- ➔ Variants **≥ 50 base pairs**
- ➔ Diversity of rearrangement types:



Inspired by Heller and Vingron (2019)

Structural Variants (SVs)

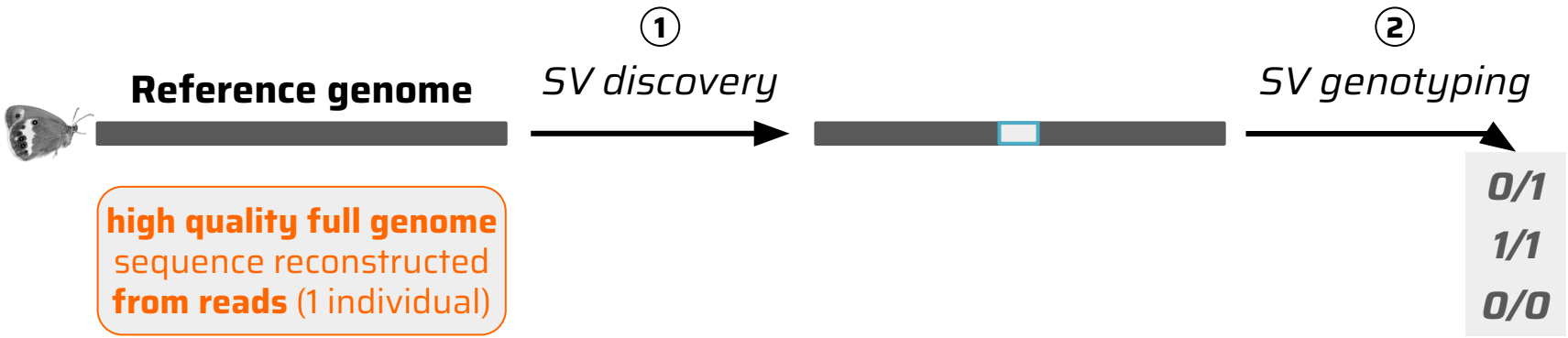
- ➔ Variants **≥ 50 base pairs**
- ➔ Diversity of rearrangement types:



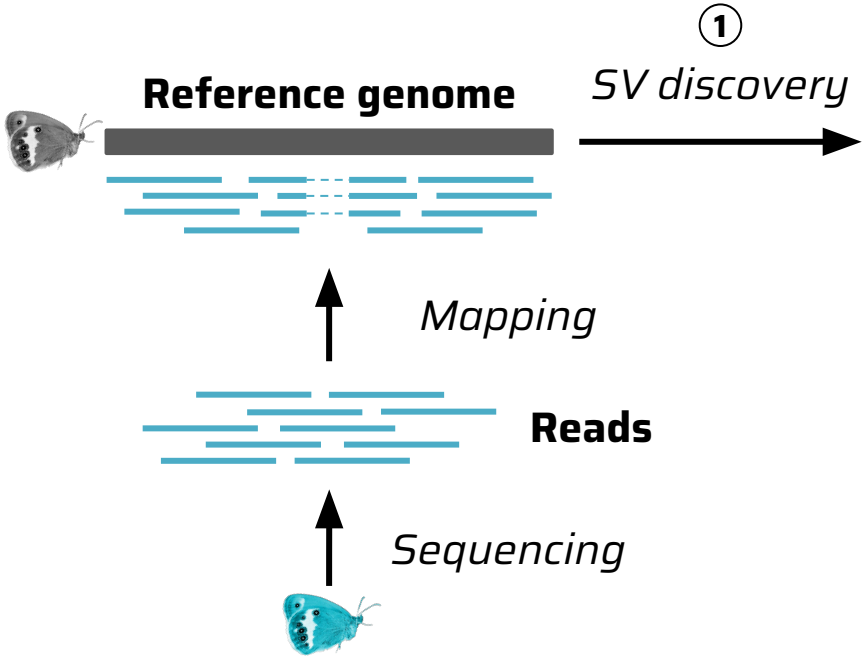
Inspired by Heller and Vingron (2019)

- ➔ **Inversions:** good candidate to explain heterogeneous gene flow along chromosomes

Classical pipeline for SV characterization



Classical pipeline for SV characterization

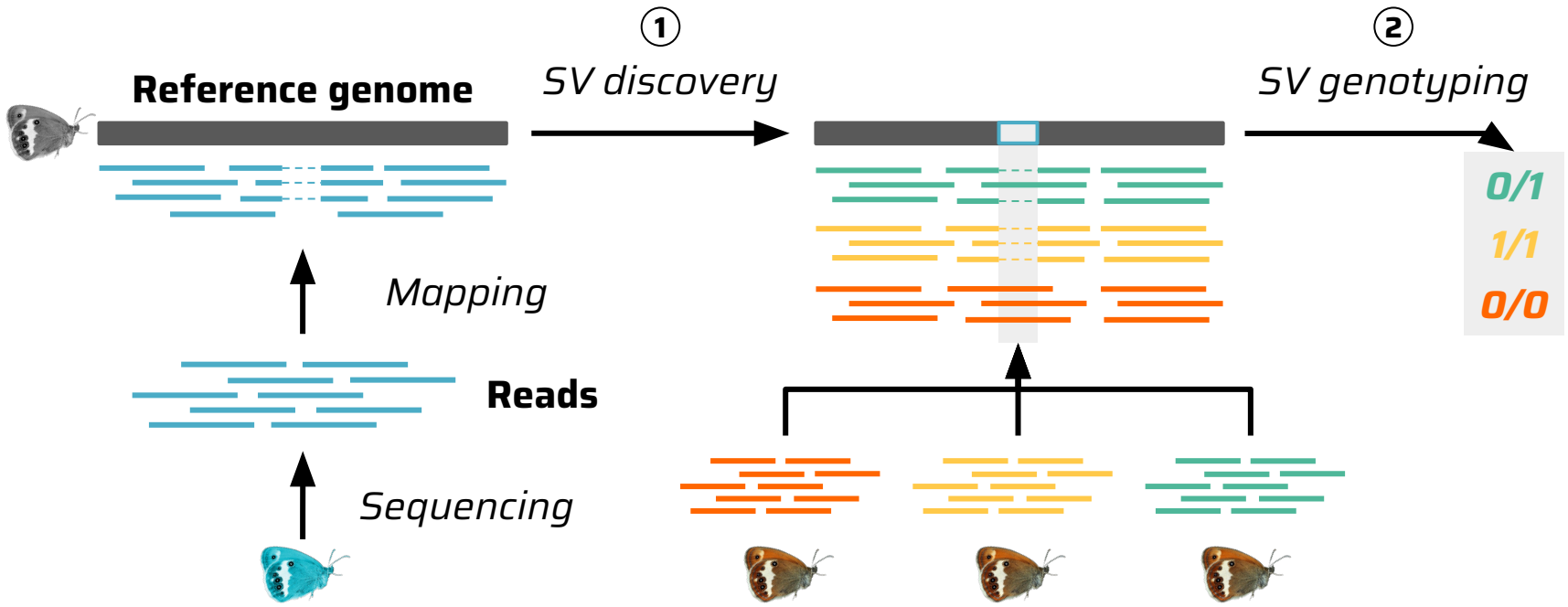



```
#CHROM  POS    INFO
ref     1000   SVTYPE=DEL;END=1250;SVSIZE=-250
ref     2500   SVTYPE=INV;END=3000
```

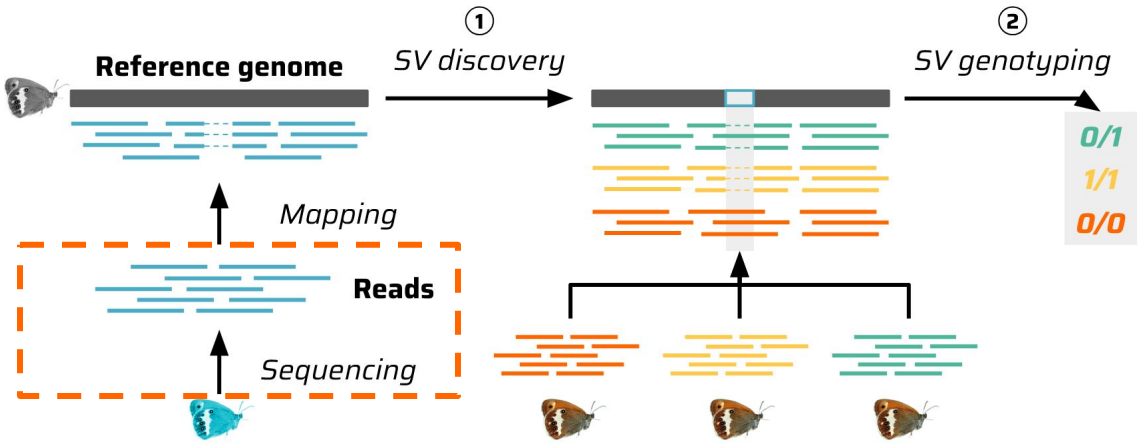
position type size

Variant
Call
Format

Classical pipeline for SV characterization



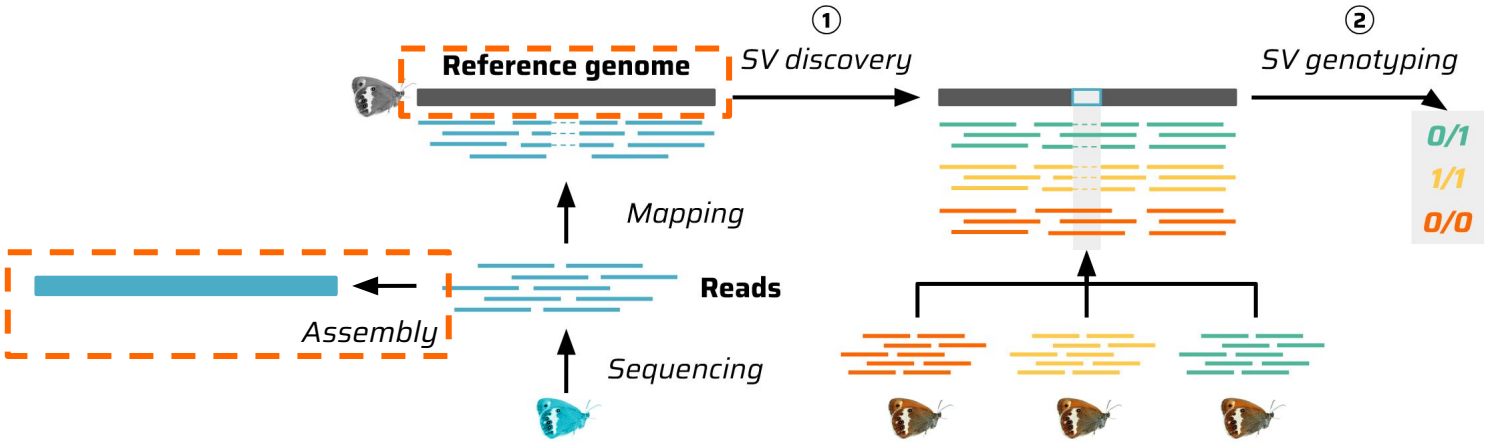
Current advances in methods for SV characterization



➔ Sequencing technologies: **accuracy and length of reads** (PacBio HiFi, Ultralong ONT)

- ***Coenonympha* sp. data:** PacBio HiFi long read sequencing for 1 individual / species

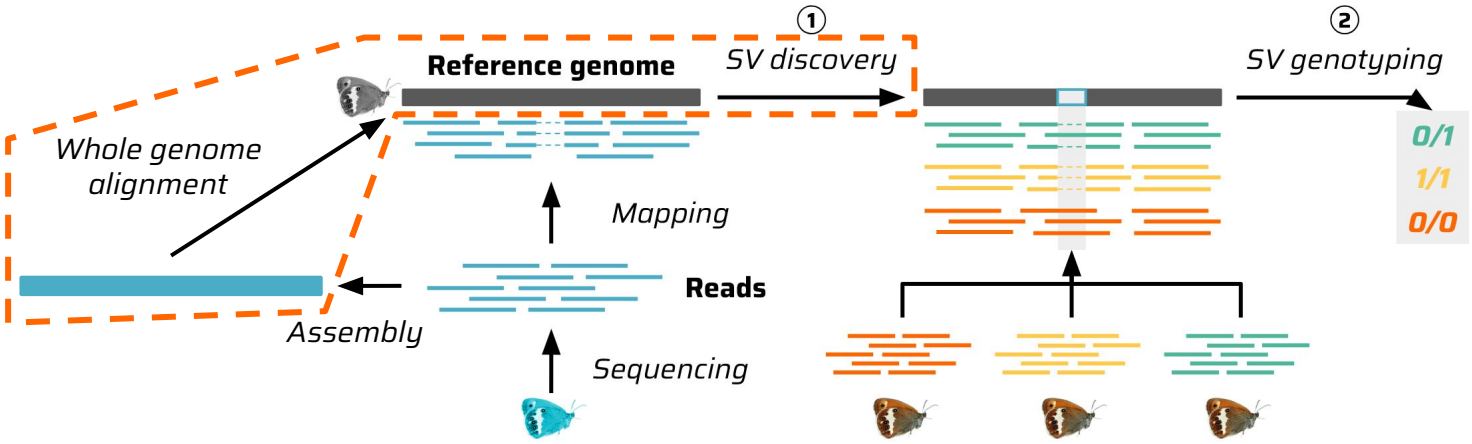
Current advances in methods for SV characterization



- ➔ Sequencing technologies: **accuracy and length of reads** (PacBio HiFi, Ultralong ONT)
- ➔ Access to **high quality genome assemblies** (model and non-model species)
  - ◆ ***Coenonympha* sp. data:** 1 assembled genome / species [1]

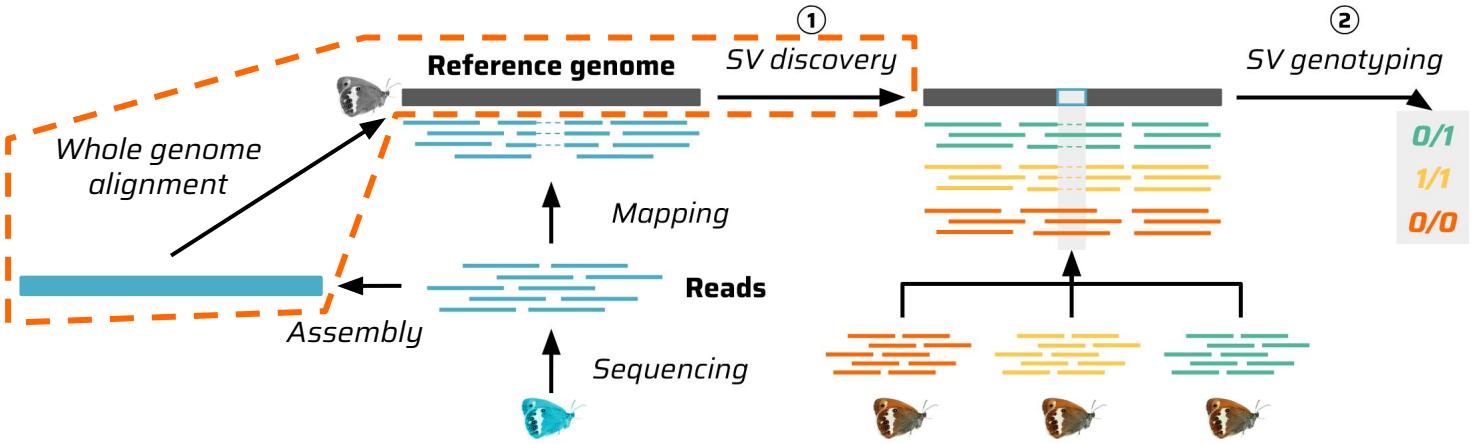
[1] Legeai, F., **Romain, S.**, Capblancq, T., Doniol-Valcroze, P., Joron, M., Lemaitre, C., & Després, L. (2024). Chromosome-level assembly and annotation of the pearly heath Coenonympha arcania butterfly genome. Genome Biology and Evolution, 16(3), evae055.

Current advances in methods for SV characterization



- ➔ Sequencing technologies: **accuracy and length of reads** (PacBio HiFi, Ultralong ONT)
- ➔ Access to **high quality genome assemblies** (model and non-model species)
- ➔ Assembly-based SV discovery ⇒ **very large SVs, divergent genomes**

Current advances in methods for SV characterization

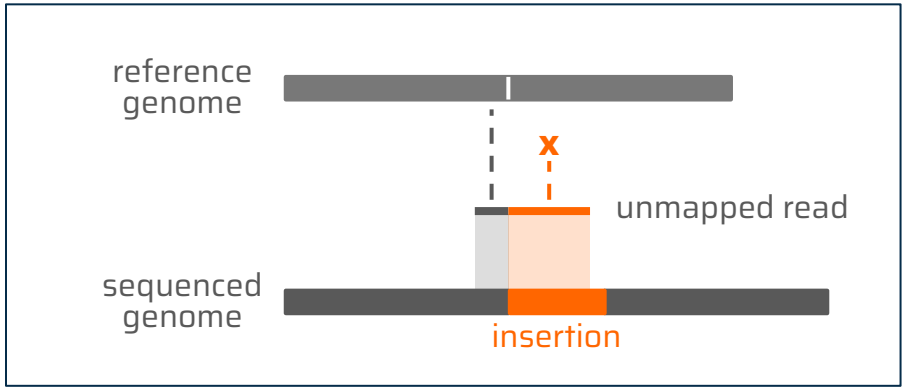


- ➔ Sequencing technologies: **accuracy and length of reads** (PacBio HiFi, Ultralong ONT)
- ➔ Access to **high quality genome assemblies** (model and non-model species)
- ➔ Assembly-based SV discovery ⇒ **very large SVs, divergent genomes**

⚠ **Reference genome and reference bias**

The issue of reference bias

➔ Reads containing non-reference alleles **less likely to be correctly mapped** than those containing reference alleles



Example of an unmapped read containing an insertion

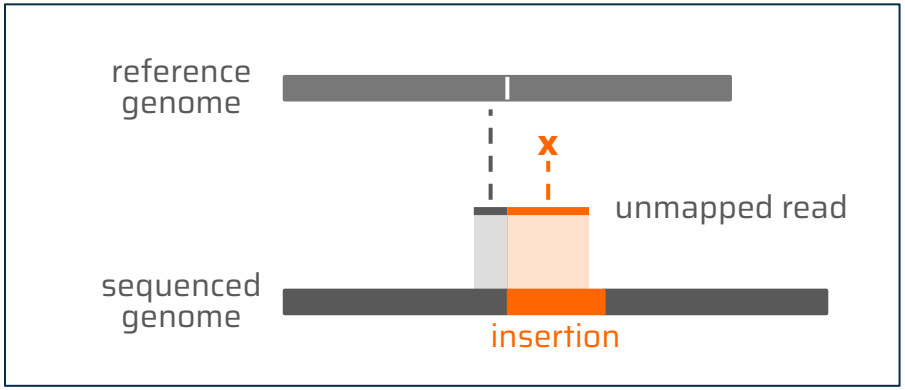
The issue of reference bias

- ➔ Reads containing non-reference alleles **less likely to be correctly mapped** than those containing reference alleles
- ➔ Occurs in:
  - ◆ **variant discovery**

    e.g. variants in non-reference sequences

  - ◆ **variant genotyping**

    bias towards reference alleles

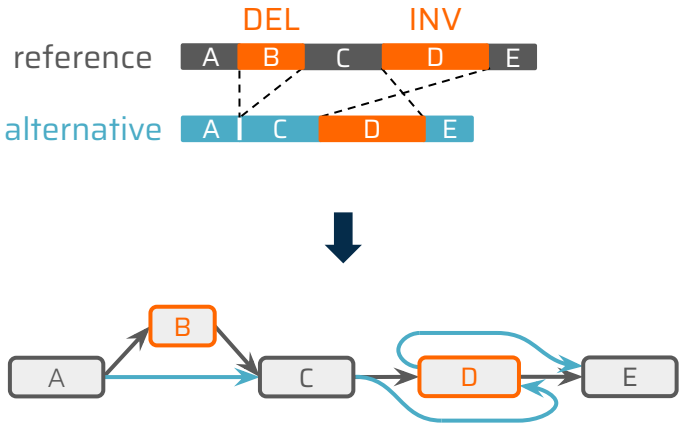


Example of an unmapped read containing an insertion

Traditional reference genome limits variant characterization.

Variation and pangenome graphs

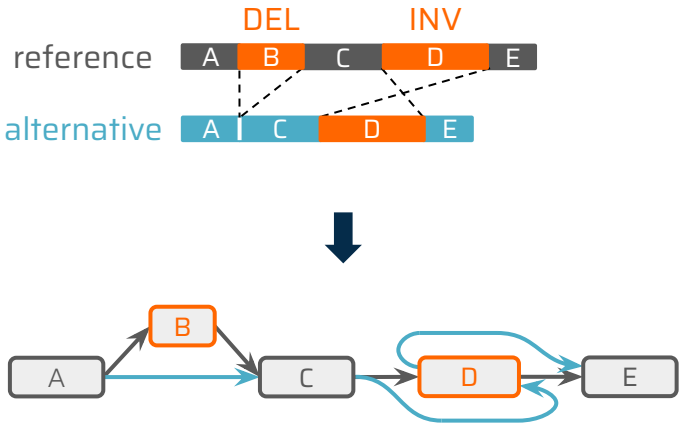
- Represent the diversity of genome sequence in populations
- Sequence graphs
  - Nodes = sequences
  - Edges = sequence adjacencies
  - Path = genome



A+,B+,C+,D+,E+ = reference path
A+,C+,D-,E+ = alternative path

Variation and pangenome graphs

- Represent the diversity of genome sequence in populations
- Sequence graphs
  - Nodes = sequences
  - Edges = sequence adjacencies
  - Path = genome
- Construction:
  - **Variation graph**: reference + variant set
  - **Pangenome graph**: whole genome alignment
- **Replace the reference genome** as base for read mapping



A+,B+,C+,D+,E+ = reference path
A+,C+,D-,E+ = alternative path

Reducing the reference bias

Objectives and contributions

To develop **new methods** for the **characterization of SVs**, in **non-model** genomes while **avoiding reference-bias.**

1. **Long-read SV genotyping on a variation graph**
2. Large inversion characterization in the *Coenonympha* alpine butterflies
3. **Investigation of inversion representation and annotation in pangenome graphs**

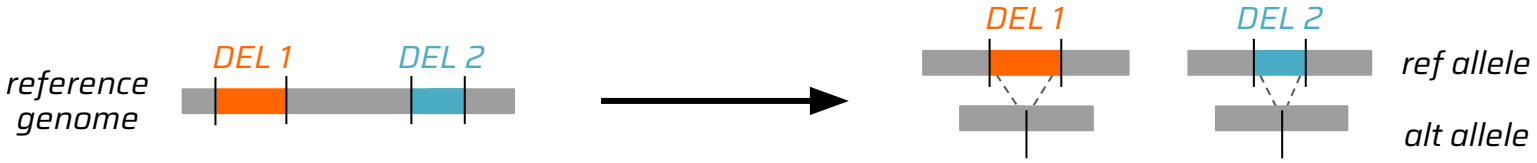
Contribution I:

Long-read structural variant genotyping on a variation graph

Limits of state of the art at the time

- **cuteSV[1], Sniffles[2], LRcaller[3], VaPoR[4]**
  - 1st step = mapping to the reference genome → **reference bias**

- **SVJedi[5]:** avoids reference bias by representing both alleles (with flanking regions)

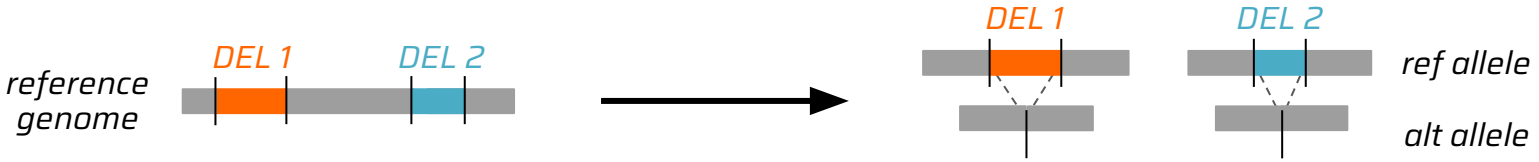


[1] Jiang *et al.* (2020), **[2]** Smolka *et al.* (2024), **[3]** Beyter *et al.* (2021), **[4]** Zhao *et al.* (2017), **[5]** Lecompte *et al.* (2020)

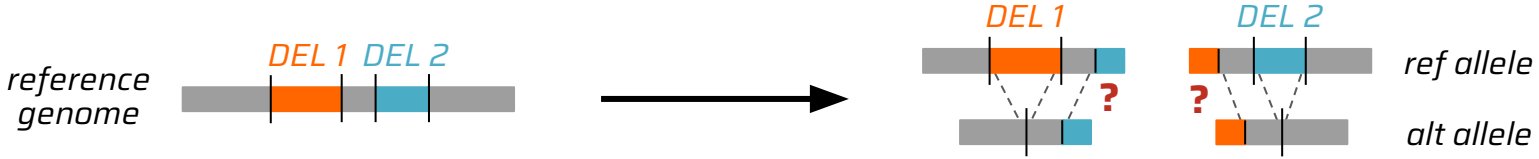
Limits of state of the art at the time

- **cuteSV[1], Sniffles[2], LRcaller[3], VaPoR[4]**
  - 1st step = mapping to the reference genome → **reference bias**

- **SVJedi[5]:** avoids reference bias by representing both alleles (with flanking regions)



- **Limit:** close and overlapping SVs

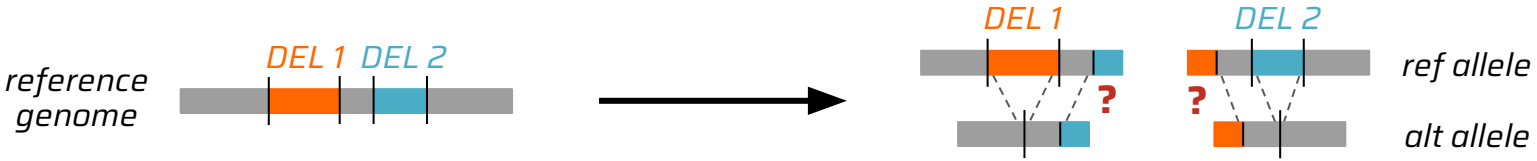


How to represent close SVs?

[1] Jiang *et al.* (2020), [2] Smolka *et al.* (2024), [3] Beyter *et al.* (2021), [4] Zhao *et al.* (2017), [5] Lecompte *et al.* (2020)

Using a variation graph to genotype SVs

➔ **SVJedi[1]:** close and overlapping SVs



➔ **Solution:** represent SVs using a **variation graph**

(**vg-Giraffe[2], Paragraph[3], GraphTyper2[4]**)

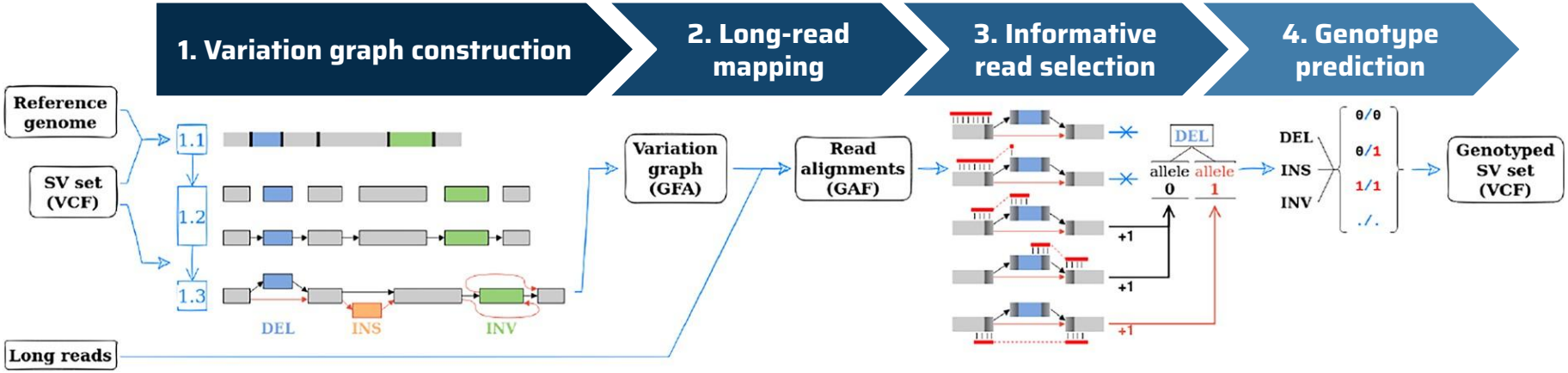
⇒ **short reads only**

No existing SV genotyping method using both long reads and variation graphs.

[1] Lecompte *et al.* (2020), **[2]** Sirén *et al.* (2020), **[3]** Chen *et al.* (2019), **[4]** Eggertson *et al.* (2019)

Development of a new method

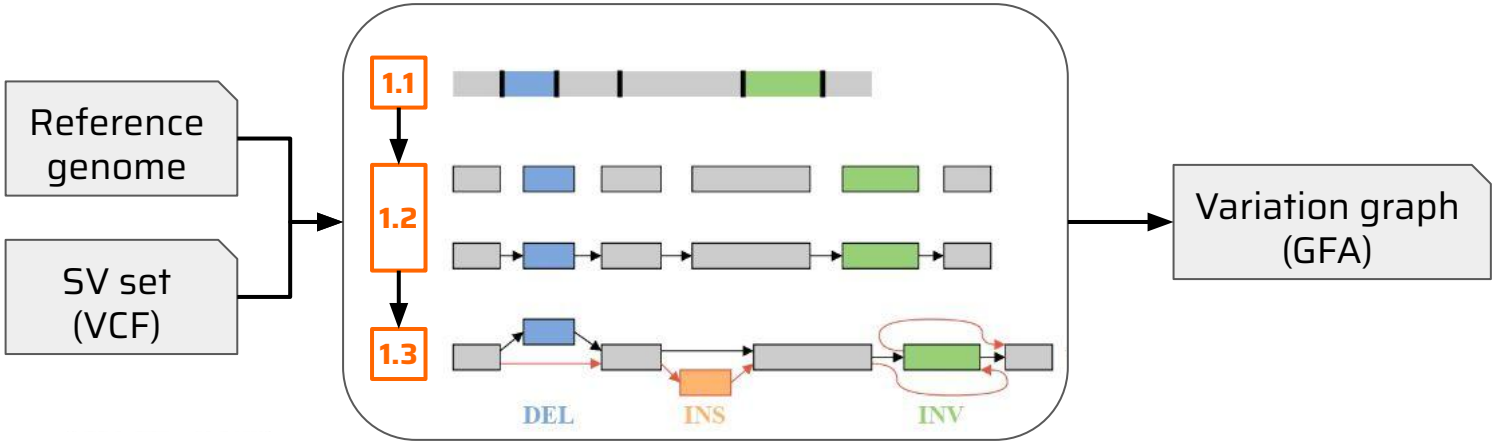
- ➔ Inspired by **SVJedi**
- ➔ **Difference:** SV representation with a **variation graph**



Development of a new method

- ➔ Inspired by **SVJedi**
- ➔ **Difference:** SV representation with a **variation graph**

Step 1: Variation graph construction



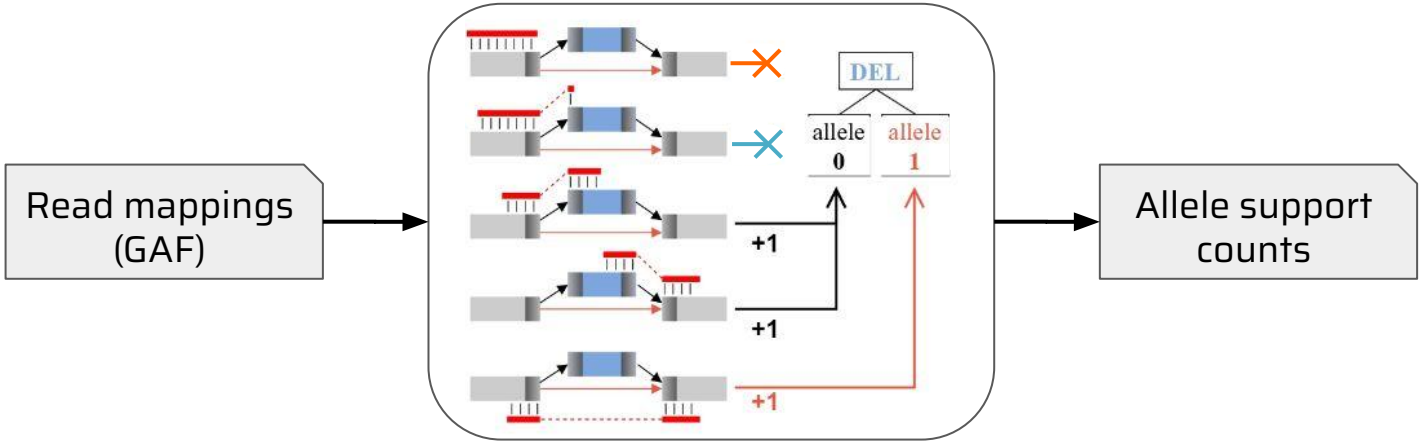
1.1 → List and sort breakpoints

1.2 → Split the reference sequence at breakpoints
→ Add a reference edges

1.3 → For each SV, add adequate alternative edge(s) + node for insertions

An edge in the graph = a SV breakpoint

Step 3: Informative read selection



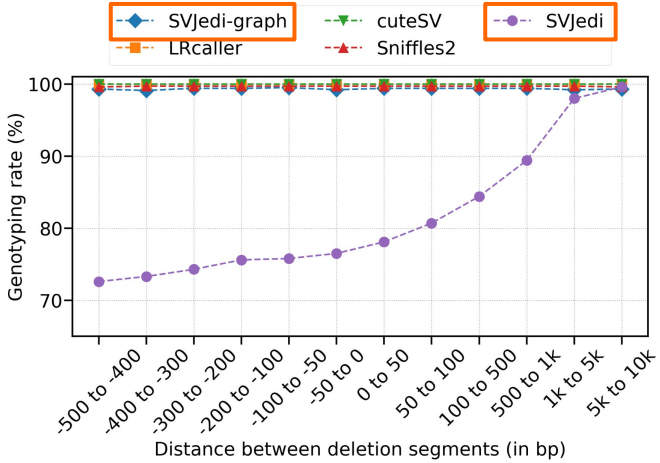
- ➔ **Filter 1:** a read must overlap a breakpoint
- ➔ **Filter 2:** a read must align over at least ***d_over*** bp before and after the breakpoint
- ➔ Increase support by 1 for each read mapping on a breakpoint

Benchmark on simulated data

➔ 12 simulated datasets on human chr 1

- 995 real deletions (dbVar)
- simulated 0/0 deletions (close to or overlapping a real deletion by ***x*** bp)

x =



Genotyping rate = % of known SV genotyped

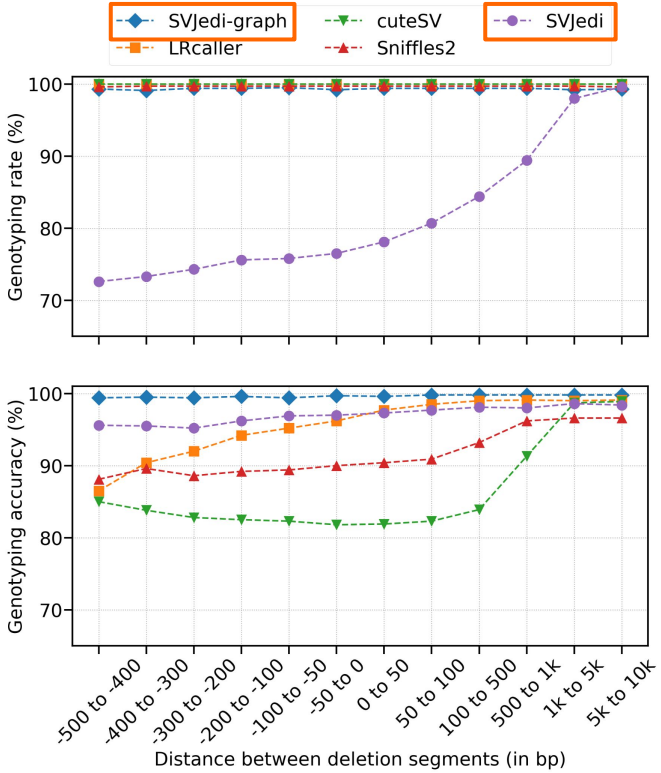
Stable genotyping rate regardless of proximity/overlapping

Benchmark on simulated data

➔ 12 simulated datasets on human chr 1

- 995 real deletions (dbVar)
- simulated 0/0 deletions (close to or overlapping a real deletion by ***x*** bp)

Genotyping accuracy = % of accurate genotypes



x =

Stable genotyping accuracy regardless of proximity/overlapping

Benchmark on human data

➔ **Genome in a Bottle** HG002 “gold standard” dataset [1]

- 12,745 SVs
- distant SVs

Long read genotyper	GiaB HG002 “gold standard”dataset	
	Accuracy (%)	Rate (%)
SVJedi-graph	**94.1**	99.5
SVJedi	86.2	**100**
cuteSV	91.3	**100**
LRcaller	89.4	99.2
Sniffles2	81.3	84.4

[1] Zook et al., Sci. Data (2016)

Benchmark on human data

➔ **Genome in a Bottle** HG002 "gold standard" dataset [1]

- 12,745 SVs
- distant SVs

➔ **Human Genome SV Consortium** HG00732 dataset [2]

- 101,743 SVs
- 56 % close SVs (< 1kb)
- 9 % overlapping SVs

Long read genotyper	GiaB HG002 "gold standard"dataset		HGSVC HG00732 dataset	
	Accuracy (%)	Rate (%)	Accuracy (%)	Rate (%)
SVJedi-graph	**94.1**	99.5	84.3	96.1
SVJedi	86.2	**100**	**89.1**	42.8
cuteSV	91.3	**100**	73.7	**100**
LRcaller	89.4	99.2	84.5	**100**
Sniffles2	81.3	84.4	80.4	99.0

Competitive on various SV set compositions.

[1] Zook et al., Sci. Data (2016), [2] Ebert et al., Science (2021)

Conclusion - Long-read SV genotyping with a VG

SVJedi-graph

- ➔ **More robust to SV closeness and faster than state of the art long-read SV genotypers**
- ➔ Implemented in Python
- ➔ Available on GitHub[1] and BioConda[2]
- ➔ Used in team collaboration projects

Long read genotyper	Running time (min) (20 CPUs)		
	Total	Mapping	Genotyping
SVJedi-graph	**29.7**	**24.8**	**4.3**
SVJedi	189.9	181.9	7.5
cuteSV	201.9	176	25.9
LRcaller	196.6	176	20.6
Sniffles2	233.9	176	57.9

Performances on the GiaB HG002 dataset with PacBio CLR reads.

- ➔ Work presented at ISMB/ECCB 2023 in Lyon
- ➔ Paper published in Bioinformatics **(Romain and Lemaitre, 2023)**

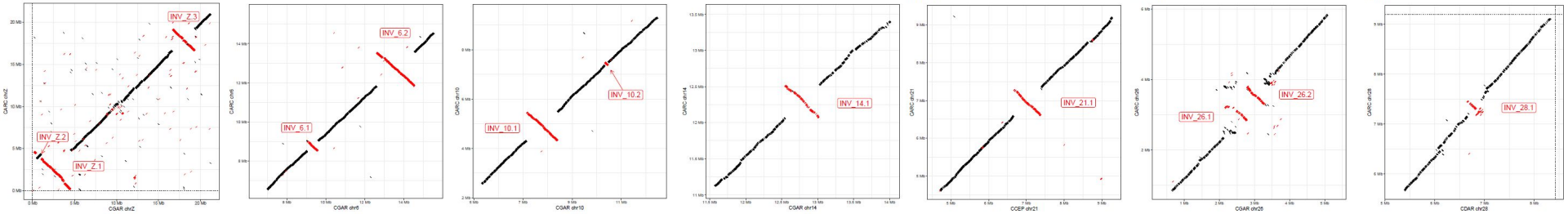
The variation graph improved long-read SV genotyping.

[1] https://github.com/SandraLouise/SVJedi-graph, [2] https://anaconda.org/bioconda/svjedi-graph

Contribution II:

Investigation of inversion representation and characterization in pangenome graphs

Large inversions in the *Coenonympha* sp. genomes



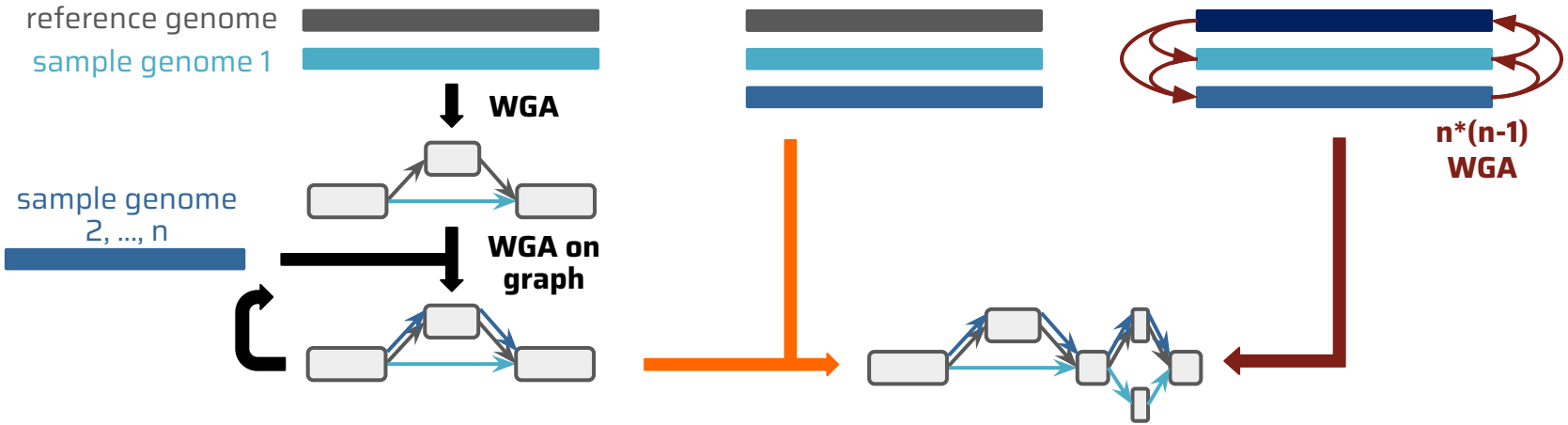
12 large inversions discovered between the *Coenonympha* sp. genomes through classical reference-based approach (WGA)

- ➔ Characterization of the *Coenonympha* inversions in a pangenome graph
  - ◆ Which pangenome graph building tool to use?
  - ◆ How to detect inversions in pangenome graphs?
- ➔ **Novelty of pangenome graphs** → inversions missing from benchmarks and analyses

How well are inversions represented and detected in pangenome graphs?

Constructing a pangenome graph

Genome set → Pangenome graph



	Minigraph [1]	Minigraph-Cactus [2]	PGGB [3]
Variants	SV	SNP, indel, SV	SNP, indel, SV
Reference	yes (backbone)	yes (backbone)	no

[1] Li *et al.* (2020), **[2]** Hickey *et al.* (2024), **[3]** Garrison *et al.* (2023)

Coenonympha sp. pangenome graphs

Genome set → Pangenome graph

- **Data:** ***Coenonympha* sp.** genomes
  - 4 diploid genomes
  - 5 (out of the 28) chromosomes

Coenonympha sp. pangenome graphs

Genome set → Pangenome graph

➔ **Data: *Coenonympha* sp.** genomes

- 4 diploid genomes
- 5 (out of the 28) chromosomes

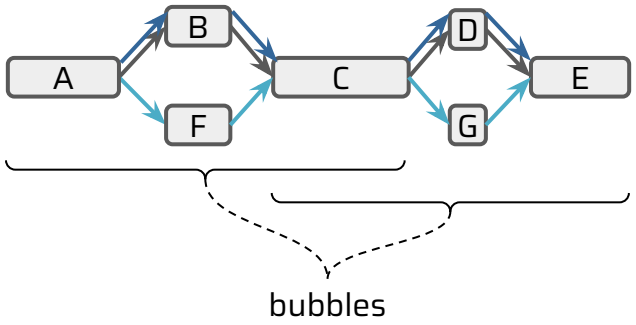
PG graph	Number of nodes (10^3)	Number of edges (10^3)
Minigraph	462	635
Minigraph-Cactus	19,667	27,246
PGGB	18,693	26,063

Pangenome graphs are complex graphs

Detection of variants in pangenome graphs

Genome set → Pangenome graph → Bubbles

➔ Bubbles = variants



Detection of variants in pangenome graphs

Genome set → Pangenome graph → Bubbles

- ➔ Bubbles = variants
- ➔ Bubble callers (*e.g.* **vg deconstruct** [1])

```
##fileformat=VCFv4.2
#CHROM    POS       ID        REF  ALT  INFO
scaffold_6  1718803   >24>26    REF  ALT  AT=>24>25>26,>24>108>26;LV=0
scaffold_6  7675607   >44>46    REF  ALT  AT=>44>45>46,>44>116>46;LV=0
```

[1] Liao *et al.* (2023)

Detection of variants in pangenome graphs

Genome set → Pangenome graph → Bubbles

- ➔ Bubbles = variants
- ➔ Bubble callers (*e.g.* **vg deconstruct** [1])

```
##fileformat=VCFv4.2
#CHROM    POS       ID        REF  ALT  INFO
scaffold_6  1718803   >24>26    REF  ALT  AT=>24>25>26,>24>108>26;LV=0
scaffold_6  7675607   >44>46    REF  ALT  AT=>44>45>46,>44>116>46;LV=0
```

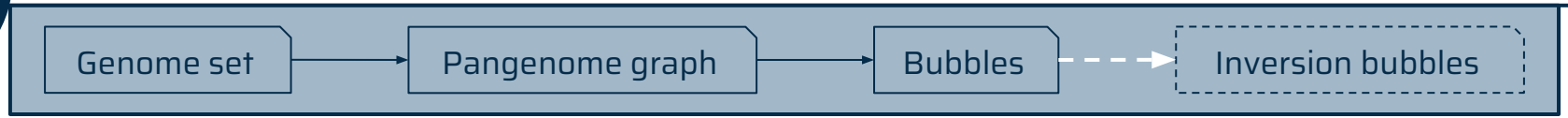
- ➔ **High number of bubbles**
- ➔ **No distinction between variant types**

PG graph	Number of bubbles
Minigraph	63,456
Minigraph-Cactus	4,750,737
PGGB	4,235,289

Need of a tool to annotate inversions among bubbles

[1] Liao *et al.* (2023)

Development of a method to annotate inversion bubbles



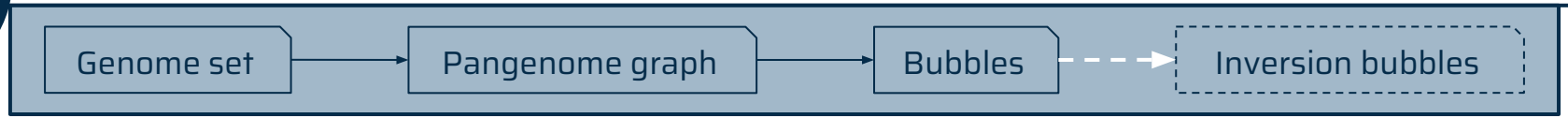
➔ **Bubble selection:** at least 2 alleles of similar size ≥ 50 bp

➔ Expected topology:

- **explicitly** represented in the **paths** of the bubble

without inner polymorphism

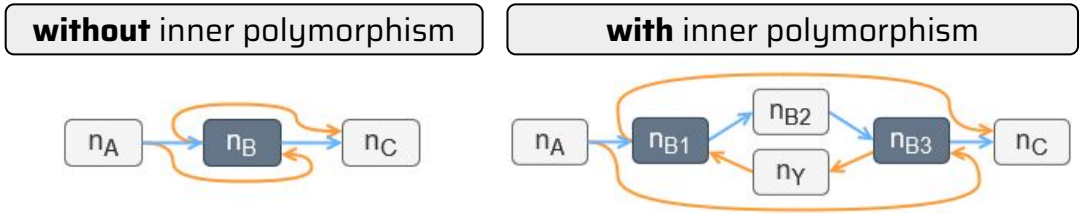
Development of a method to annotate inversion bubbles



➔ **Bubble selection:** at least 2 alleles of similar size ≥ 50 bp

➔ Expected topology:

- **explicitly** represented in the **paths** of the bubble



Development of a method to annotate inversion bubbles

Genome set → Pangenome graph → Bubbles → Inversion bubbles

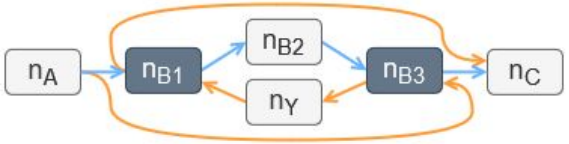
➔ **Bubble selection:** at least 2 alleles of similar size ≥ 50 bp

➔ Expected topology:

- **explicitly** represented in the **paths** of the bubble

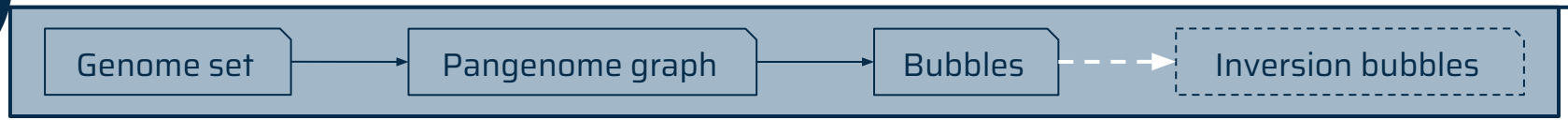
without inner polymorphism

with inner polymorphism



Another topology for inversions in the graphs?

Development of a method to annotate inversion bubbles



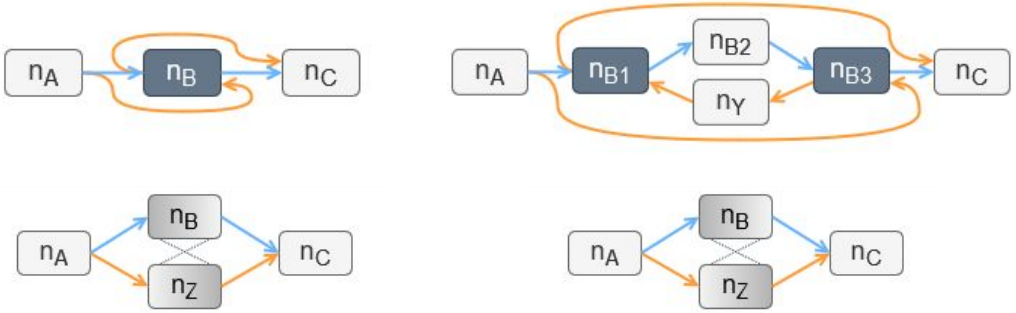
➔ **Bubble selection:** at least 2 alleles of similar size ≥ 50 bp

➔ Expected topologies:

- **explicitly** represented in the **paths** of the bubble
- represented as **unrelated sequences**

rescued through **alignment**

without inner polymorphism | **with** inner polymorphism



INVPG-annot: looking for inversion signal in both allele **paths** and **alignments**

Final annotation in *Coenonympha* sp. pangenomes

Genome set → Pangenome graph → Bubbles → Inversion bubbles

➔ **Data: *Coenonympha* sp.** genomes

- 4 diploid genomes
- 5 chromosomes
- **8 known inversions (≥ 100 kb)**

PG graph	Number of known inversions annotated
Minigraph	0
Minigraph-Cactus	**2** (path-explicit)
PGGB	**2** (path-explicit)

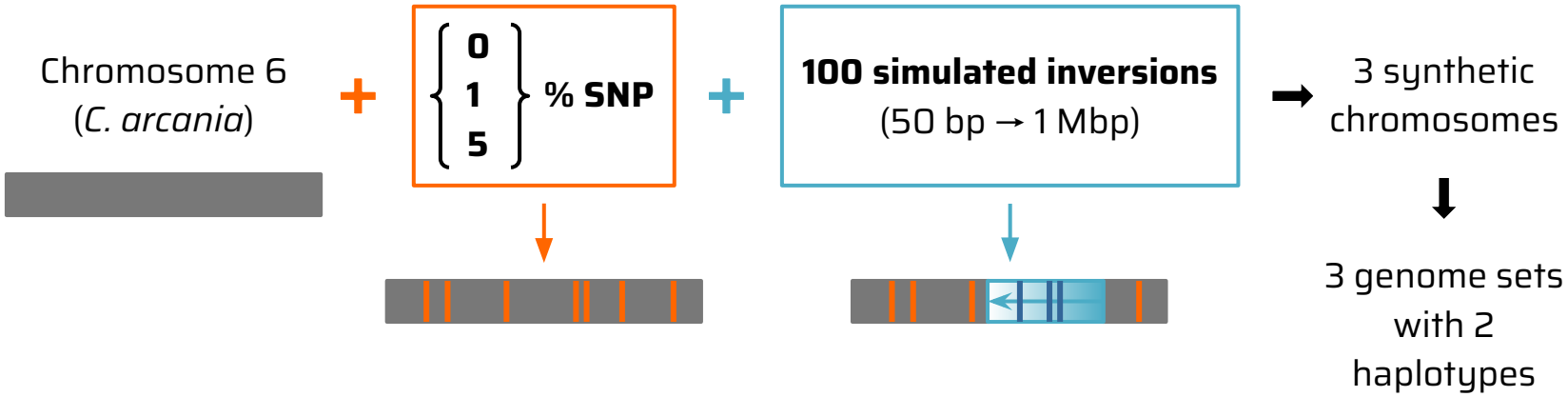
Why are so many inversions missed?

Investigating with simulated datasets

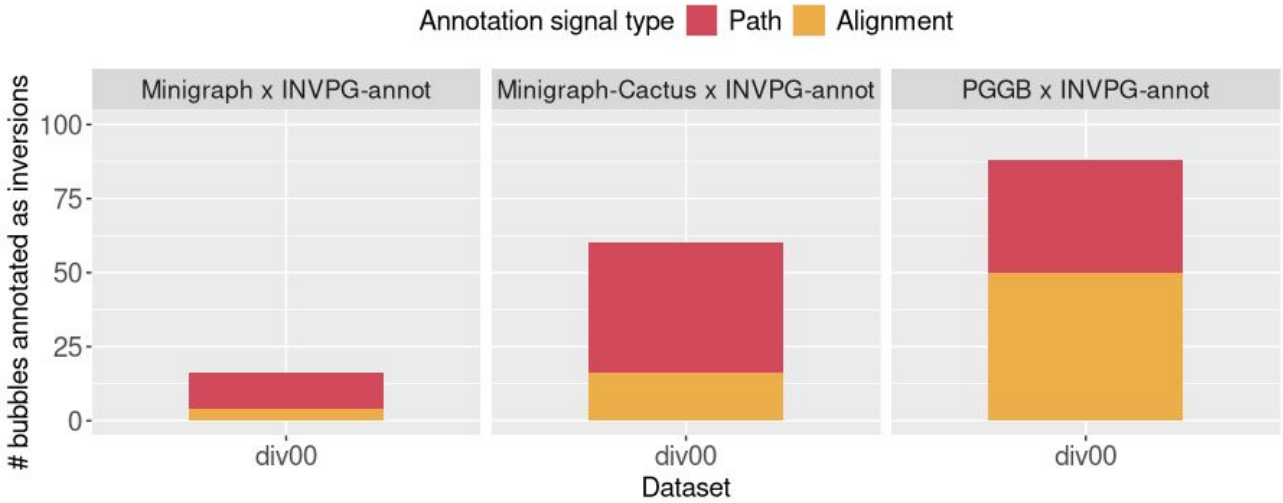
Genome set → Pangenome graph → Bubbles → Inversion bubbles

➔ Hypotheses

- **Genome divergence** hinders the representation of inversions in pangenome graphs
- **Inversion size** influence their topology in pangenome graphs

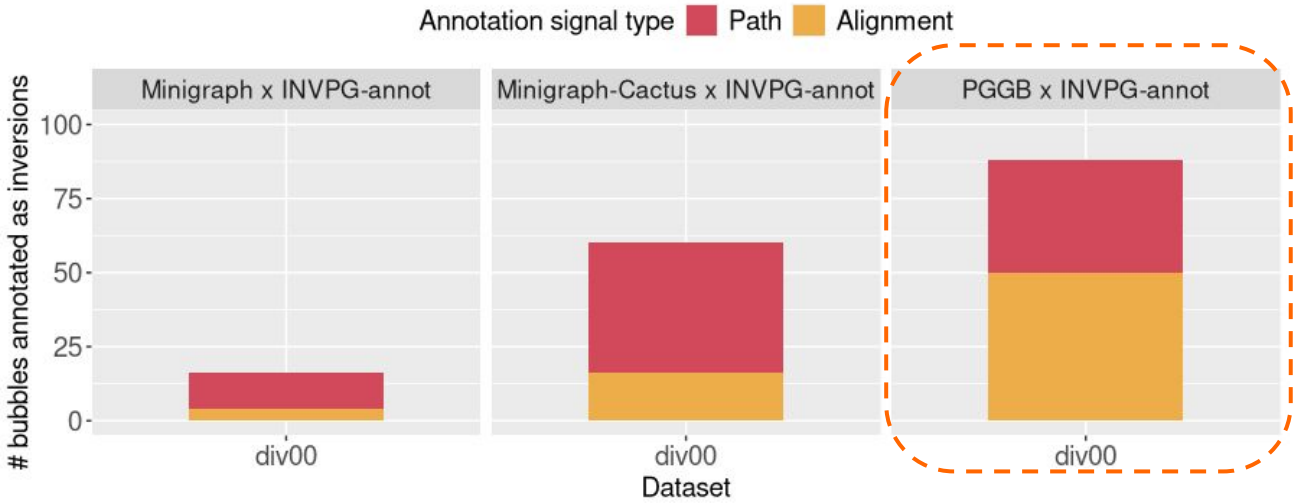


Results - Simplest case without punctual polymorphism



➔ Most inversions not represented explicitly

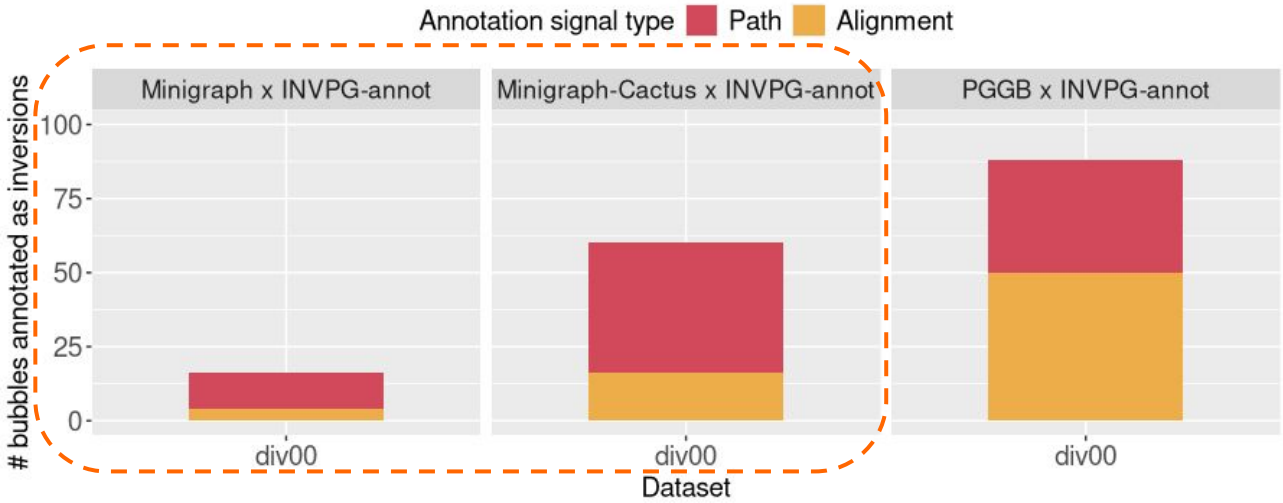
Results - Simplest case without punctual polymorphism



➔ Most inversions not represented explicitly

- many **alignment-rescued** (**PGGB**)

Results - Simplest case without punctual polymorphism

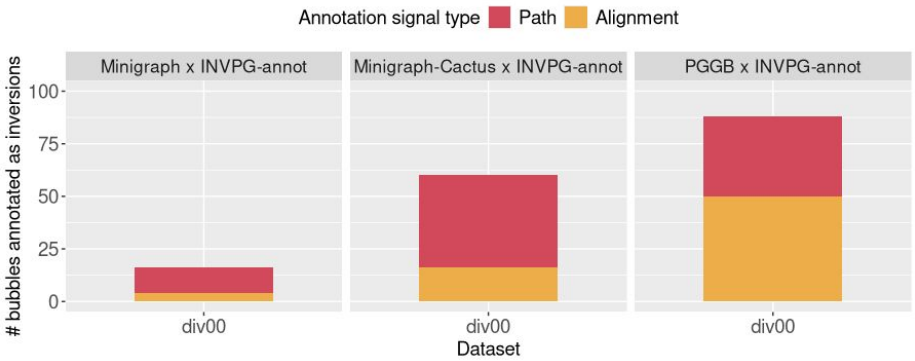


➔ Most inversions not represented explicitly

- many **alignment-rescued** (**PGGB**)
- many **not annotated** (**Minigraph**, **Minigraph-Cactus**)

Results - Simplest case without punctual polymorphism

➔ Most inversions not represented explicitly



Recall = % of true inversions corresponding to at least one annotated bubble

➔ **Low recall:** Minigraph, Minigraph-Cactus

PG tool	Recall (%)
Minigraph	**5**
Minigraph-Cactus	**10**
PGGB	86

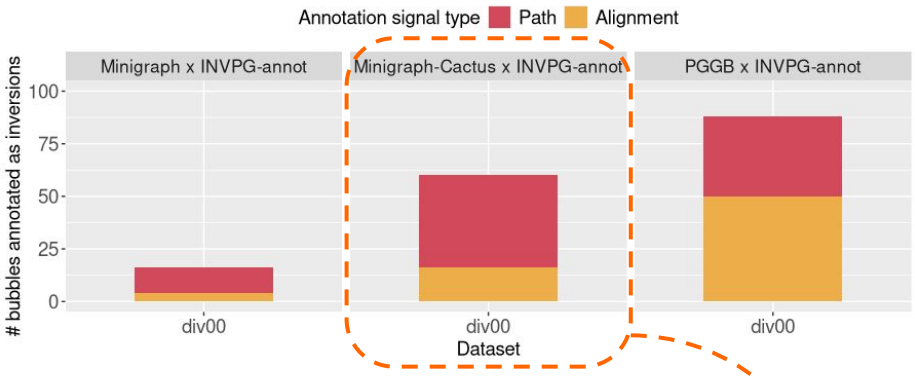
Results - Simplest case without punctual polymorphism

➔ Most inversions not represented explicitly

Recall = % of true inversions corresponding to at least one annotated bubble

➔ **Low recall:** Minigraph, Minigraph-Cactus

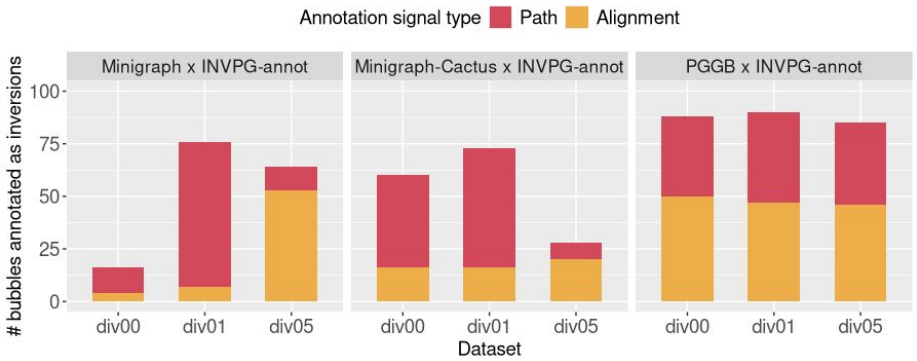
➔ **Redundancy** (no false positives)



PG tool	Recall (%)
Minigraph	**5**
Minigraph-Cactus	**10**
PGGB	86

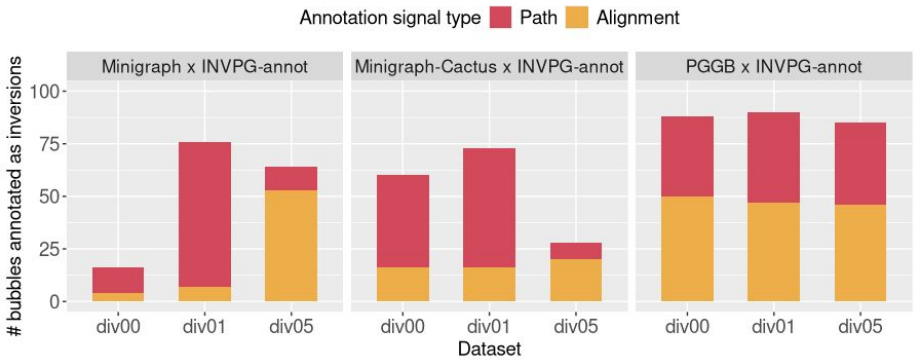
Results - Increasing the genomic divergence

- High divergence (5 %):
  - More **alignment-rescued** inversions (**Minigraph**)
  - Less annotated inversions (**Minigraph-Cactus**)



Results - Increasing the genomic divergence

- ➔ High divergence (5 %):
  - ◆ More **alignment-rescued** inversions (**Minigraph**)
  - ◆ Less annotated inversions (**Minigraph-Cactus**)

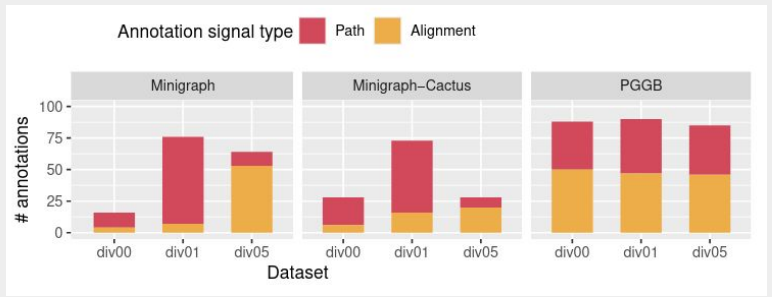


- ➔ Best recall: **1%** divergence (all tools)
- ➔ **Minigraph-Cactus least tolerant** to higher divergence

PG tool	**Recall (%)**		
	Div 0%	**Div 1%**	**Div 5%**
Minigraph	5	**65**	59
Minigraph-Cactus	8	**55**	27
PGGB	86	**88**	83

Conclusion - Investigation of inversions in PG graphs

- **Representation of inversions in pangenome graphs varies with the constructing tool.**
  - **Genome divergence:** impacts inversion annotation in Minigraph-Cactus graphs.
  - **Inversion size:** determines the topology in PGGB graphs.



Number and topology of annotated inversions with simulated data

- **INVPG-annot:** inversion annotation tool implemented in Python, available on GitHub[1]
- Work presented at **EAGS** in Toulouse, 2024 (**E**nvironmental and **A**gronomical **G**enomics **S**ymposium)
- **Still in progress** (follow up: 0.1% divergence, test other bubble callers)

[1] https://github.com/SandraLouise/INVPG_annot

General Conclusion and Perspectives

Contributions

➔ Long-read SV genotyping using a variation graph

- **SVJedi-graph** [https://github.com/SandraLouise/SVJedi-graph]
- paper published in Bioinformatics **(Romain S. and Lemaitre C., 2023)**

➔ Discovery of 12 large inversions in the alpine *Coenonympha* species

- paper submitted soon to Genome Biol Evol **(Romain S., Capblancq T., Desprès L., Joron M., Legeai F., Lemaitre C., *in prep.*)**

➔ Exploration of inversion representation in pangenome graphs

- **INVPG-annot** [https://github.com/SandraLouise/INVPG_annot]
- paper in preparation **(Romain S., Dubois S., Legeai F. and Lemaitre C., *in prep.*)**

Inversion discovery: inter-specific application

➔ **Reference-based strategy**

- ➕ more well tested tool options
- ➕ most complete large inversion detection in *Coenonympha*
- ➖ aligners tailored to intra-specific divergence

➔ **Perspective**

- ◆ introduce more choice for aligners

➔ **Pangenome-based strategy**

- ➕ simultaneous comparison of many genomes
- ➖ inversions often not represented explicitly
- ➖ a lot of missed inversions in real data

➔ **Perspective**

- ◆ bypass the intermediary bubble detection tools → directly from the graph

Perspectives: Long-read SV genotyping using graphs

- **SVJedi-graph** tested on *Coenonympha* sp. inversions
  - Sensitiveness to **breakpoint imprecision** and genome **divergence**
  - **Improvement:** analysis of **split mappings**

- Adaptation of the method to **pangenome graphs**?
  - **Advantages:**
    - higher read mapping rate for multi-species SV characterization
    - more and more commonly produced
  - **Drawbacks:**
    - more complex graphs than variation graphs (# nodes/edges)
    - not only SVs in the graph (SNPs, indels)
    - needs precise characterization of SVs

Acknowledgments

Jury members

- Birte Kehr
- François Sabot
- Sèverine Berard
- Pierre Peterlongo

CSI members

- Camille Marchet
- Mathieu Rousseau-Gueutin

PhD supervisors

- Claire Lemaitre
- Fabrice Legeai

Divalps collaborators

in particular

- Thibaut Capblancq
- Laurence Desprès
- Mathieu Joron

As well as all colleagues from **GenScale** (special thanks to Siegfried Dubois), **Dyliss** and the **GenOuest** platform.

Université de Rennes · cnrs · Inria · UMR IRISA · GENSCALE · anr

References

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., ... & Stefansson, K. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nature genetics, 53(6), 779-786.

Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., ... & Eberle, M. A. (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome biology, 20, 1-13.

Dabbaghie, F., Ebler, J., & Marschall, T. (2022). BubbleGun: enumerating bubbles and superbubbles in genome graphs. Bioinformatics, 38(17), 4217-4219.

Eggertsson, H. P., Kristmundsdottir, S., Beyter, D., Jonsson, H., Skuladottir, A., Hardarson, M. T., ... & Melsted, P. (2019). GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nature communications, 10(1), 5402.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., ... & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature biotechnology, 36(9), 875-879.

Garrison, E., Guarracino, A., Heumos, S., Villani, F., Bao, Z., Tattini, L., ... & Prins, P. (2023). Building pangenome graphs. Biorxiv, 2023-04.

Heller, D., & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. Bioinformatics, 35(17), 2907-2915.

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., ... & Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. Nature biotechnology, 42(4), 663-673.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., ... & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome biology, 21, 1-24.

References

Lecompte, L., Peterlongo, P., Lavenier, D., & Lemaitre, C. (2020). SVJedi: genotyping structural variations with long reads. Bioinformatics, 36(17), 4568-4575.

Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. Genome biology, 21, 1-19.

Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., ... & Paten, B. (2023). A draft human pangenome reference. Nature, 617(7960), 312-324.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. Genome biology, 20, 1-14.

Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. Genome biology, 21(1), 253.

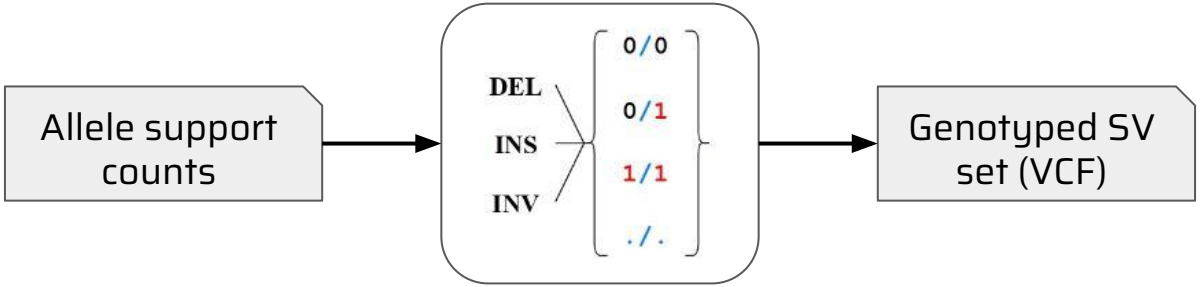
Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., ... & Paten, B. (2020). Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. BioRxiv, 2020-12.

Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., ... & Sedlazeck, F. J. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nature biotechnology, 1-10.

Zhao, X., Weber, A. M., & Mills, R. E. (2017). A recurrence-based approach for validating structural variation using long-read sequencing technology. Gigascience, 6(8), gix061.

Appendix - SVJedi-graph's method - Genotyping model

1. VG construction | 2. Long-read mapping | 3. Read support count | **4. Genotype prediction**



➔ Genotype likelihood
- bi-allelic SVs
- diploid genome

$$\mathscr{L}(0/0) = (1-p_e)^{c_0} \times p_e^{c_1} \times C_{c_0+c_1}^{c_0}$$

$$\mathscr{L}(1/1) = p_e^{c_0} \times (1-p_e)^{c_1} \times C_{c_0+c_1}^{c_0}$$

$$\mathscr{L}(0/1) = \left(\frac{1}{2}\right)^{c_0+c_1} \times C_{c_0+c_1}^{c_0}$$

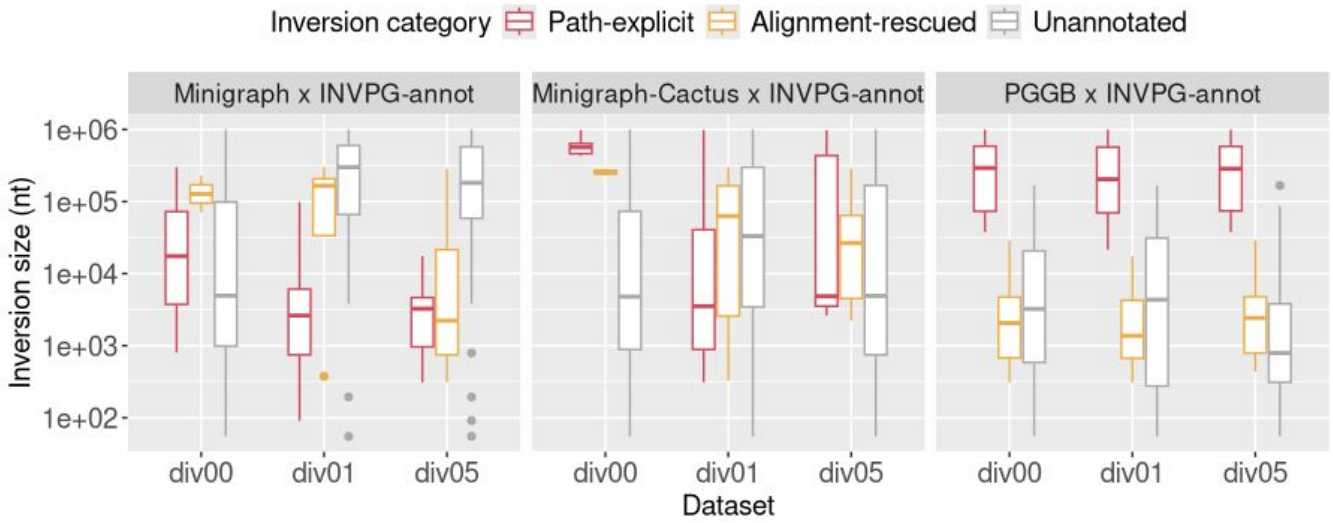
c_0 = # reads supporting allele 0
c_1 = # reads supporting allele 1
p_e = probability of a read to map erroneously on an allele

Appendix - Inversions in PG graphs - Simulated datasets

Genome set	PG tool	Graph size (Mb)	# Nodes (10^3)	# Edges (10^3)	# Bubbles	Recall (%)
0% SNP	Minigraph	25.49	0.1	0.2	48	5
	Minigraph-Cactus	27.38	204.0	220.5	16,201	10
	PGGB	20.30	3.1	4.3	5,114	86
1% SNP	Minigraph	21.43	0.3	0.3	107	65
	Minigraph-Cactus	24.02	864.1	1,144.1	279,837	55
	PGGB	20.18	587.6	783.5	195,771	88
5% SNP	Minigraph	21.39	0.3	0.3	89	59
	Minigraph-Cactus	22.32	2,318.1	3,073.5	755,405	27
	PGGB	21.65	2,803.7	3,740.6	945,302	83

Appendix - Inversions in PG graphs - Inversion size

➔ **PGGB:** inversion topology driven by inversion size



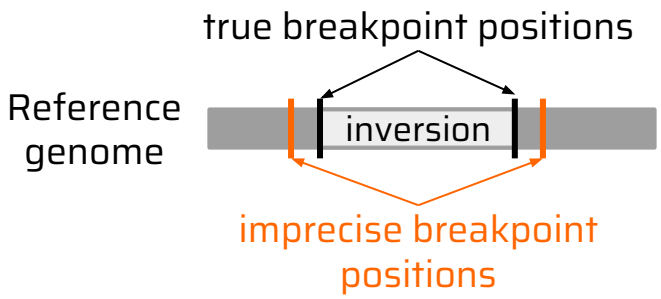
Appendix - Inversions in PG graphs - Haplotype number

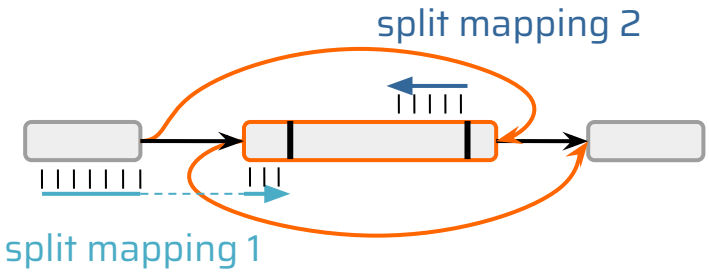
➔ Increase of annotation recall with 10 haplotypes (at divergence 0%)

PG graph	**Recall (%)**	
	2 haplotypes	**10 haplotypes**
Minigraph	5	11
Minigraph-Cactus	10	20
PGGB	86	90

Appendix - SVJedi-graph's perspectives - Split mappings

➔ Analyze **split mappings** to reduce sensitiveness to **breakpoint imprecision**





Currently: split mappings analyzed independently
→ erroneous or no genotype