Alignment-free de novo structural variations from pangenomes

Who am I?

PhD in Computer Science (2016/2019)
University of Milano-Bicocca
Prof. Paola BONIZZONI

Post-doc at BIAS (2020)
University of Milano-Bicocca
Prof. Paola BONIZZONI

Post-doc at SeqBio (2020/2022)
Institut Pasteur, Paris
Dr. Rayan CHIKHI

Post-doc at BIAS (2022/2024)
University of Milano-Bicocca
Prof. Paola BONIZZONI

Post-doc at CompBio (from Sep.2024)
Comenius University in Bratislava
Prof. Tomáš VINAŘ

Project Information

ASVA-CGR
Grant agreement ID: 101180581

DOI
10.3030/101180581

EC signature date
25 June 2024

Start date 1 September 2024 — **End date** 31 August 2026

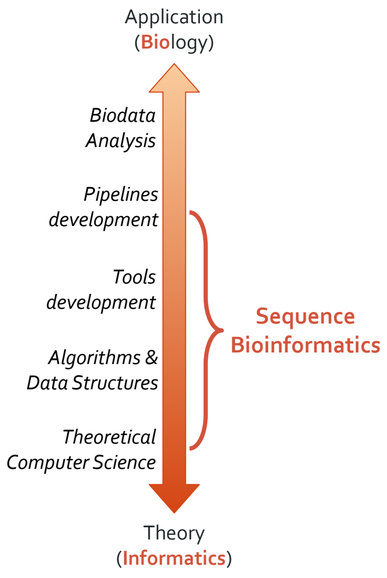
Funded under
Widening participation and spreading excellence

Total cost
No data

EU contribution
€ 149 219,52

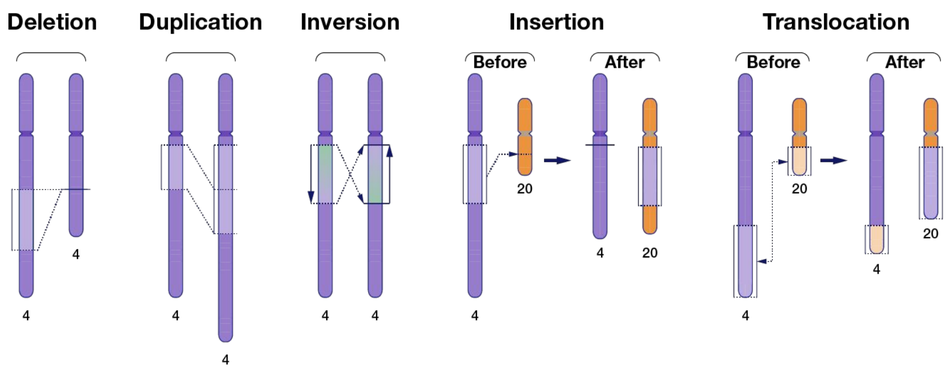
Coordinated by
UNIVERZITA KOMENSKEHO V BRATISLAVE
Slovakia

"Algorithms for Structural Variation Analysis in Challenging Genomic Regions"



Context

Structural Variation: big (≥50bp) chromosomal rearrangements



SVs account for **genetic disorders** *(autism, schizophrenia…)* and **cancer** *(melanoma, breast, prostate…)*

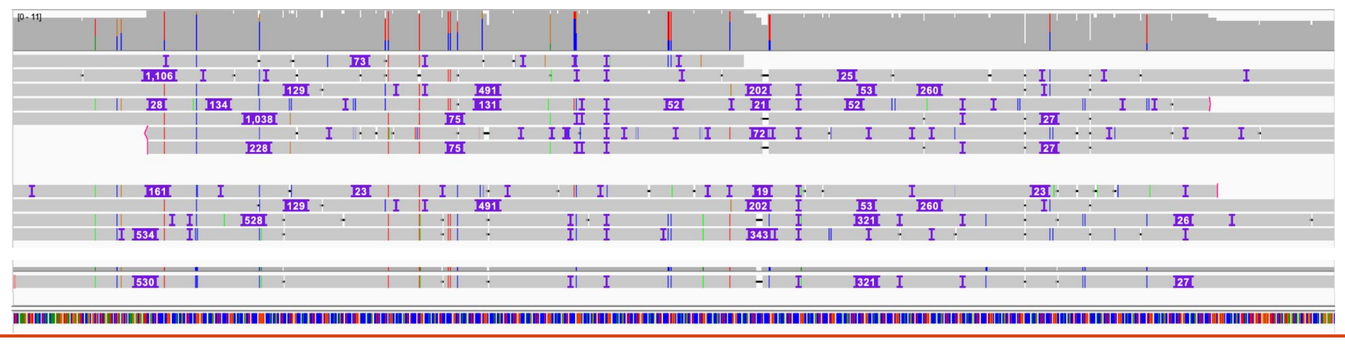
Image from https://www.genome.gov

State-of-the-art and CGR

A lot of ***reference-based*** approaches
- **short reads alignment (✗ inaccurate)**
  - …
- **long reads alignment (✗ "inaccurate")**
  - *SVIM [Bioinformatics, 2019]*
  - *cuteSV [Genome Biology, 2020]*
  - *SVDSS [Nature Methods, 2023]*
  - *sniffles2 [Nature Biotechnology, 2024]*
  - …
- **genome assembly (✗ expensive)**
  - *dipcall [Nature Methods, 2018]*
  - *PAV [Science, 2021]*
  - *VolcanoSV [Nature Communication, 2024]*
  - …

Few **pangenome-based** approaches **(✗ expensive)** *[bioRxiv, 2024]*

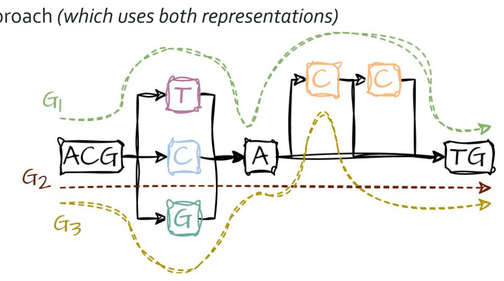
But what about Challenging Genomic Regions?



Proposal

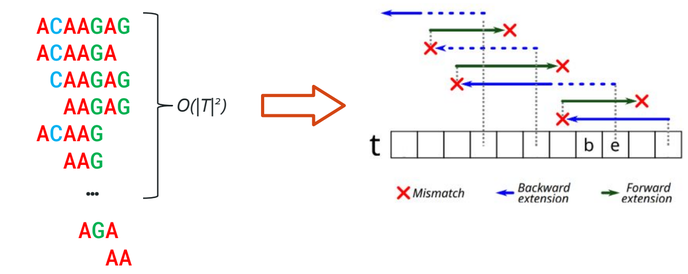
✓ **pangenome-based** approach *(which uses both representations)*

G_1 = ACGTACCTG
G_2 = ACGCATG
G_3 = ACGGACTG



✓ based on efficient FMD index (BWT-based) and **specific strings**

R = { ACATGAG }
T = { ACAAGAG }



✓ **mapping-free** approach *(does not rely on read alignments)*

Methods

(i) index pangenome

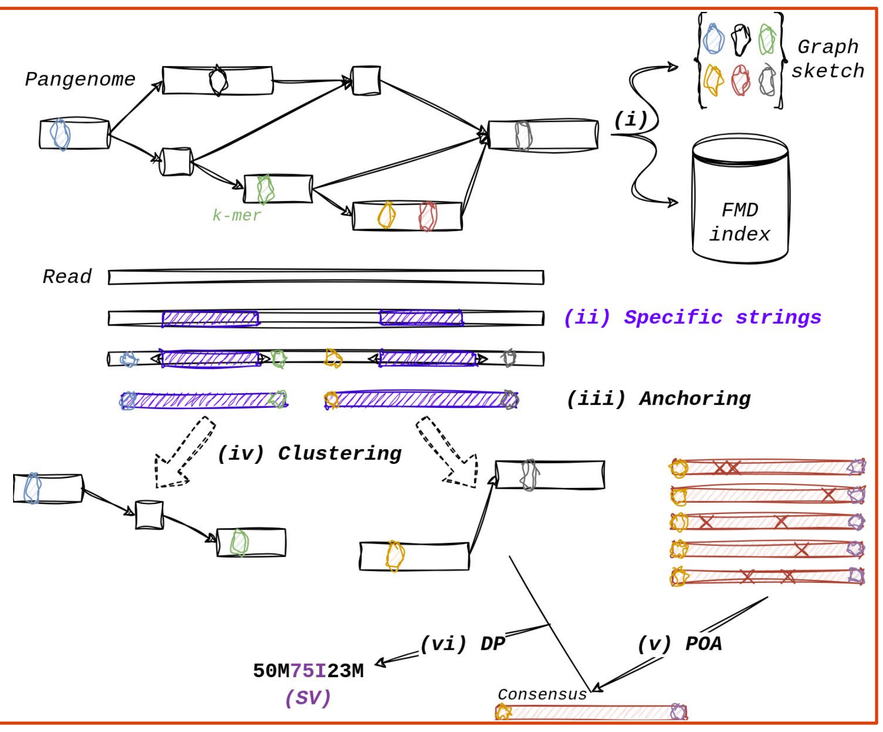
(ii) compute specific strings using the FMD-index

(iii) anchor specific strings to graph using the k-mer based sketches

(iv) assign clustering anchor specific strings to reduce graph complexity

(v) summarize information coming from different reads over the same locus

(vi) realign consensus back to local graph and get structural variation



Results & WIP

Currently testing a C++ prototype on chr21 from T2T, HPRC variant sets, and simulated PacBio HiFi data

Tool	Precision	Recall	F1
pansv	67.18	55.53	60.80
dipcall	80.46	60.88	69.32
sniffles2	69.07	54.59	60.99
cuteSV	72.23	51.46	60.09

→ improve approach
- enhance anchoring step
- replace POA with ad-hoc methodology that can exploit anchors

→ perform extensive experimental evaluation
- full-genome with genome stratification
- real data *(both PacBio and ONT)*

→ write the paper

https://github.com/ldenti/pansv

This work is supported by the European Union's Horizon programme under grant agreement No. 101180581 (ASVA-CGR)

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
Univerzita Komenského v Bratislave

MATFYZ CONNECTIONS

Luca Denti, Tomáš Vinař, Broňa Brejová
Comenius University Bratislava, Slovakia

Fereydoun Hormozdiari
UC Davis, California (US)

Paola Bonizzoni
University of Milano-Bicocca, Italy