# Computational PANGenomics 2022

## #CPANG22

Instituto Gulbenkian de Ciência, Portugal
Day 3 - 2022/05/25

Erik Garrison and Andrea Guarracino

# Vertebrate pangenome graphs are complex

A major challenge is writing software that can deal with graphs representing **hundreds of eukaryotic genomes**.

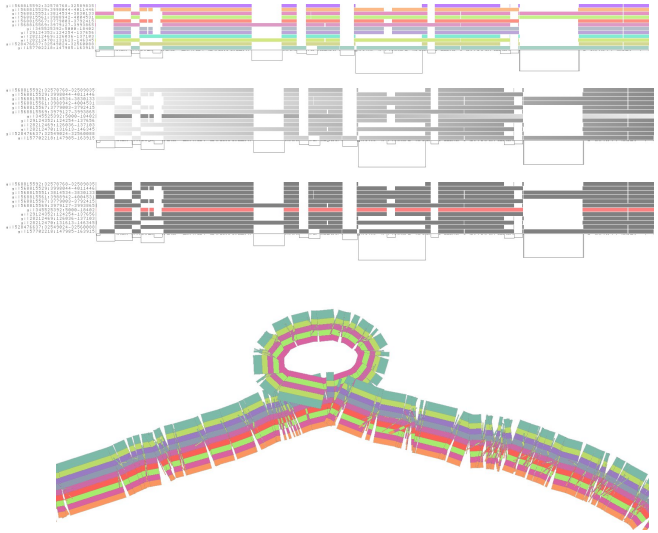
**Highly repetitive regions** (centromeres, segmental duplications, and acrocentric chromosomes) increases the complexity of the operations performed on graphs.

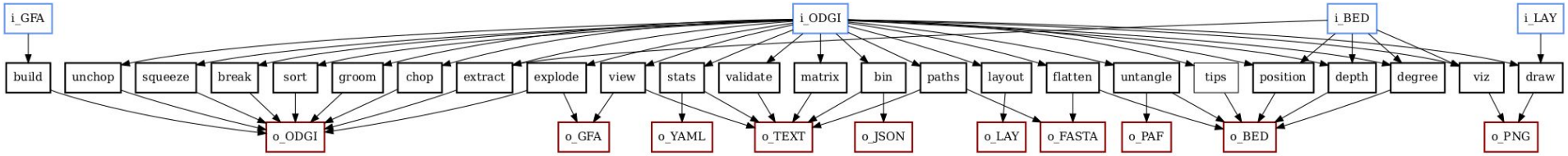
Centromeric region of a chr1 pangenome graph made with 44 human *de novo* assemblies from the Human Pangenome Reference Consortium (HPRC) dataset. Figure made with odgi draw.

# Our solution: a new suite of tools for pangenome graphs

To overcome these problems, we have developed an **Optimized Dynamic Genome/Graph Implementation** (ODGI), a new suite of tools to work with pangenome graphs structured in the variation graph model.

ODGI offers more than 30 tools for graph interrogation, manipulation, and visualization.



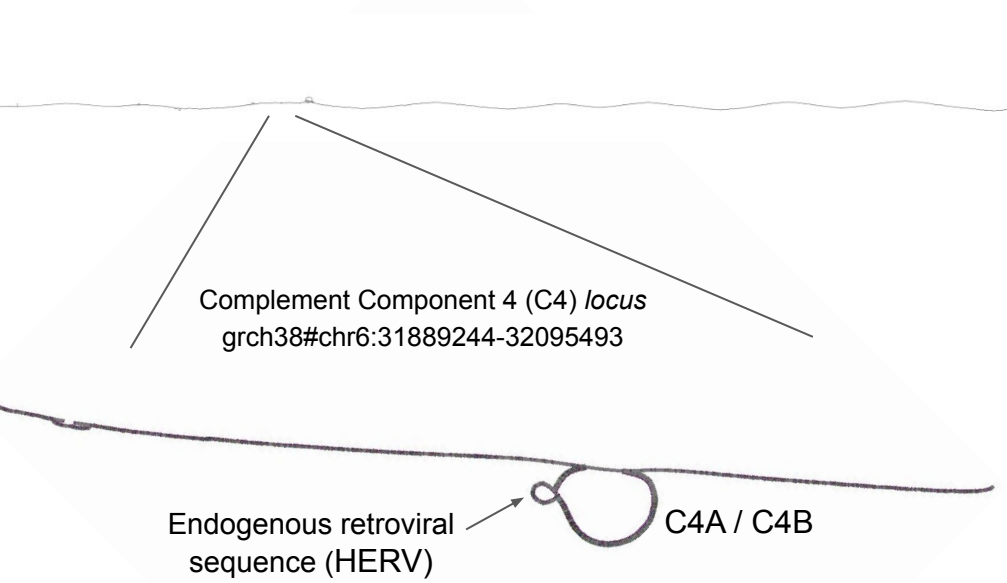


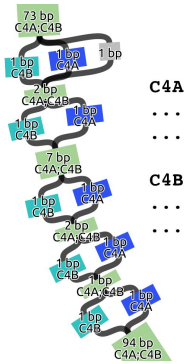

Methods provided by ODGI (in black) and their supported input (in blue) and output (in red) data formats.

# Dissecting pangenome graphs - `odgi extract`

## Downstream analyses may require focusing on specific *loci* in the pangenome.

Pangenome graph of the human chromosome 6 with 90 haplotypes (44 diploid *de novo* assemblies plus the GRCh38 and CHM13 reference genomes). A portion of the 2D layout is shown.




**C4A**

```
... ACC CCC TGT CCA GTG TTA GAC AGG ...
... Thr Pro Cys Pro Val Leu Asp Arg
```

**C4B**

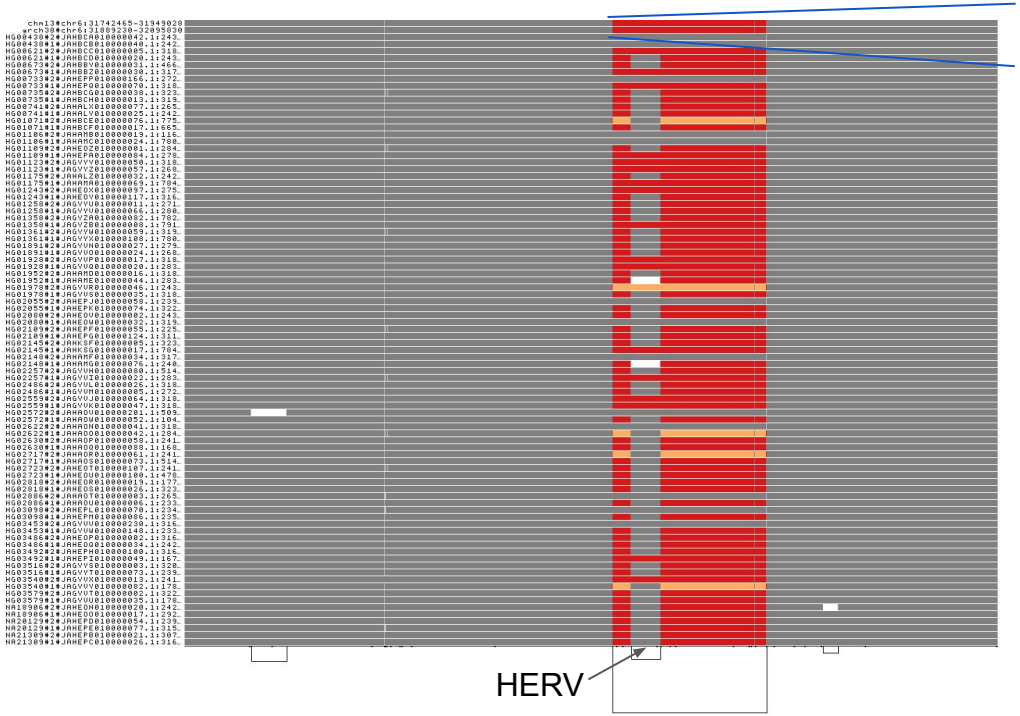
```
... ACC CTC TCT CCA GTG ATA CAT AGG ...
... Thr Leu Ser Pro Val Ile His Arg ...
```

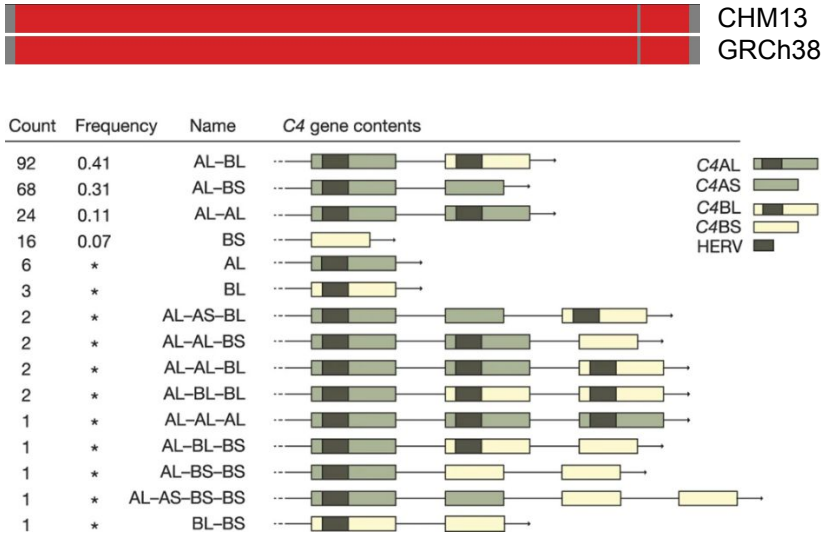
C4-copy determining residues (Sekar et al., 2016).
Figure made with Bandage and odgi position.

# Dissecting pangenome graphs - `odgi extract`

Pangenome graph of the C4 *locus* with 90 haplotypes (44 diploid *de novo* assemblies plus the GRCh38 and CHM13 reference genomes).

Colored by path depth (white = 0x, grey ~ 1x, red ~ 2x, yellow ~ 3x)

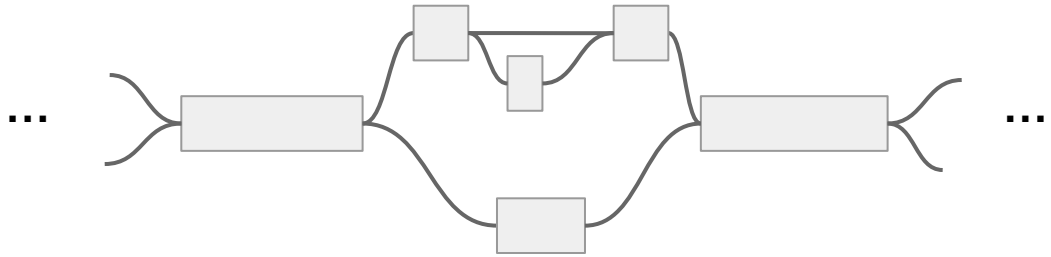



222 chromosomes from 111 genetically unrelated individuals. Figure from [Sekar et al., 2016](Sekar et al., 2016).

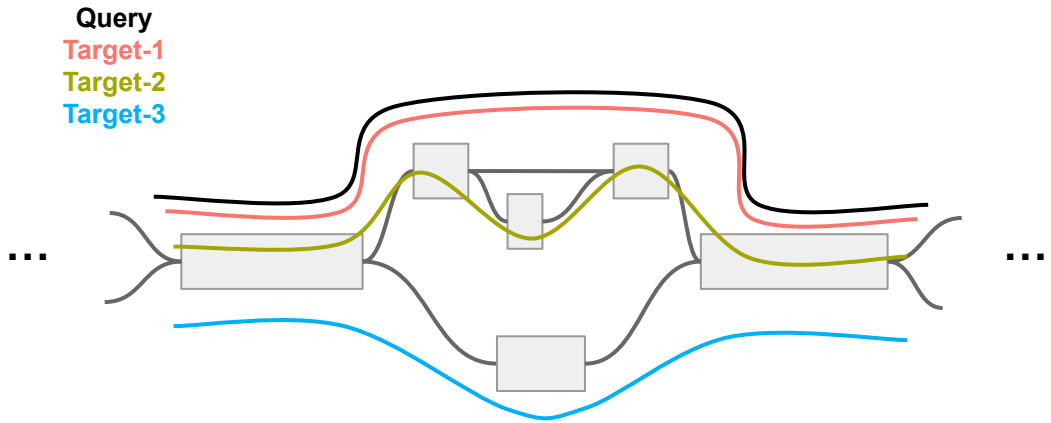
# Pangenome untangling

*Untangling* extracts pairwise alignments from variation graphs.



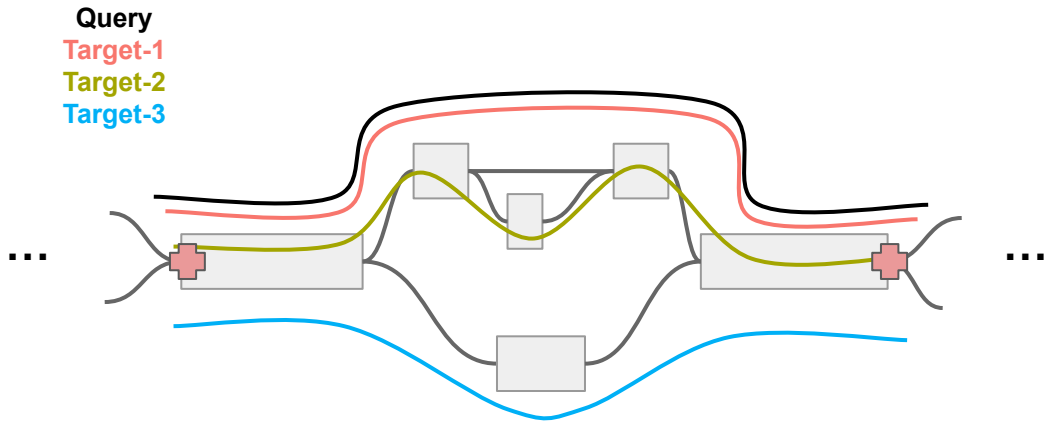
# Pangenome untangling

*Untangling* extracts pairwise alignments from variation graphs.



# Pangenome untangling

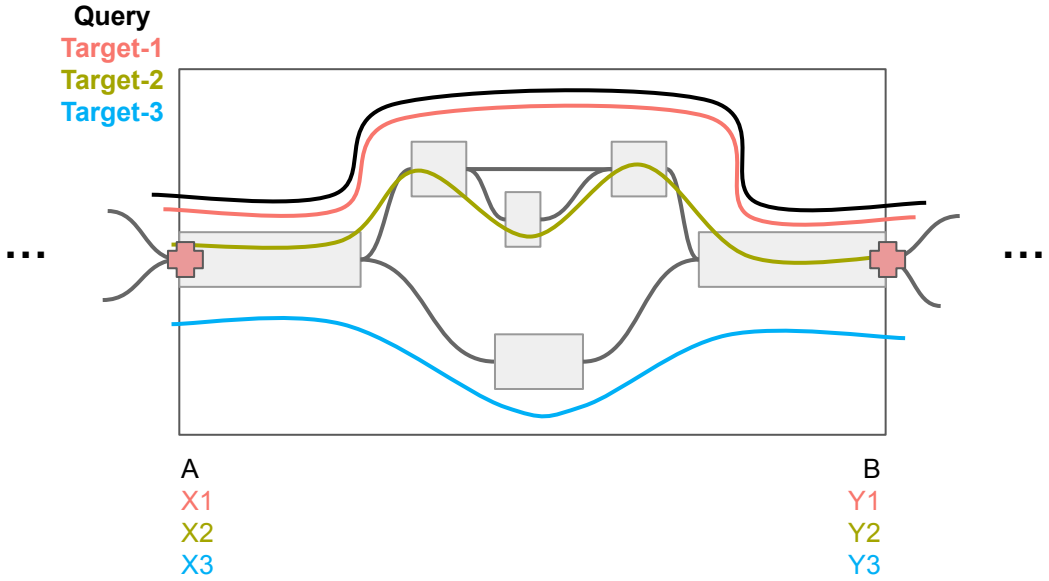
*Untangling* extracts pairwise alignments from variation graphs.




Identify cut points in the graph

# Pangenome untangling

*Untangling* extracts pairwise alignments from variation graphs.

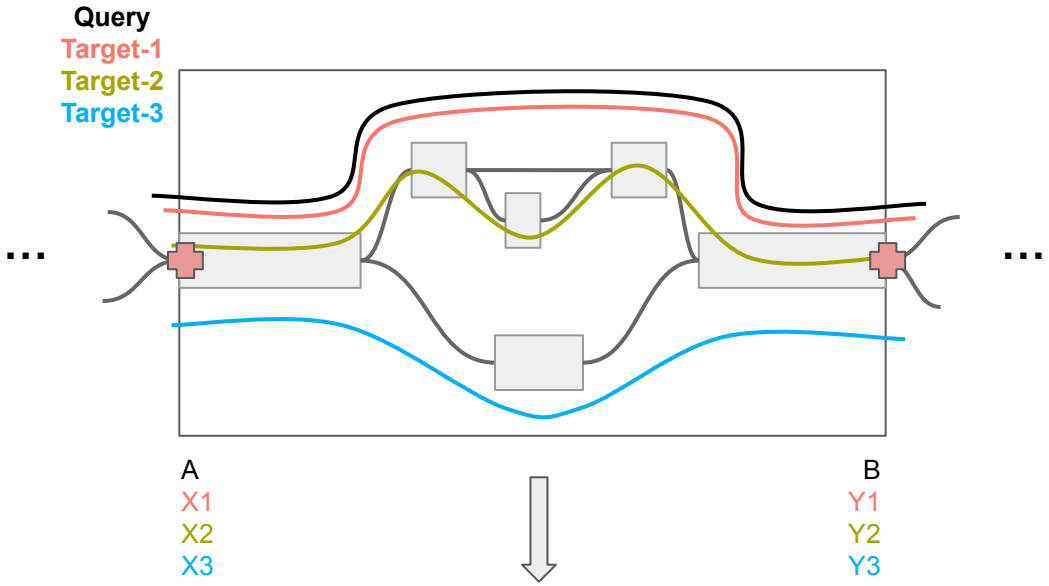



Identify cut points in the graph

Define segment boundaries

# Pangenome untangling

*Untangling* extracts pairwise alignments from variation graphs.




| query | start | end | target | start | end | jaccard | rank |
|---|---|---|---|---|---|---|---|
| Query | A | B | Target-1 | X1 | Y1 | 1 | 1 |
| Query | A | B | Target-2 | X2 | Y2 | 0.95 | 2 |
| Query | A | B | Target-3 | X3 | Y3 | 0.7 | 3 |

Identify cut points in the graph

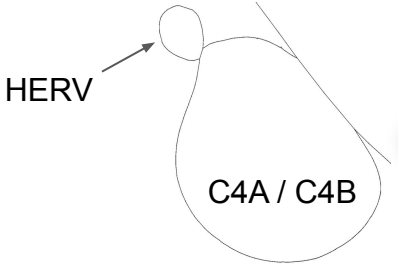
Define segment boundaries

Compare segments

with jaccard in node (allele) space

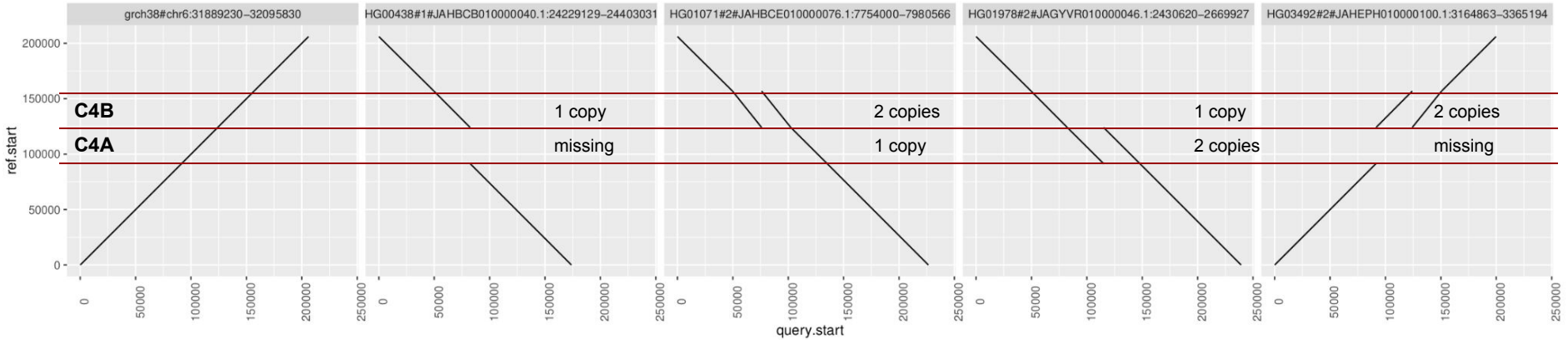
# Untangling pangenome graphs - `odgi untangle`

Repetitive sequences produce collapsed repeats in the pangenome graphs.




Paired BED output

| query.name | query.start | query.end | ref.name | ref.start | ref.end | score | inv | self.cov | nth.best |
|---|---|---|---|---|---|---|---|---|---|
| HG03492#2#JAHEPH010000100.1:3164863-3365194 | 0 | 91469 | chm13#chr6:31742465-31949028 | 0 | 91441 | 0.998121 | + | 1 | 1 |
| HG03492#2#JAHEPH010000100.1:3164863-3365194 | 91469 | 124145 | chm13#chr6:31742465-31949028 | 124001 | 156853 | 0.99149 | + | 1.80362 | 1 |
| HG03492#2#JAHEPH010000100.1:3164863-3365194 | 124145 | 150633 | chm13#chr6:31742465-31949028 | 124001 | 156853 | 0.802825 | + | 1.99185 | 1 |
| HG03492#2#JAHEPH010000100.1:3164863-3365194 | 150633 | 199828 | chm13#chr6:31742465-31949028 | 156853 | 206060 | 0.997848 | + | 1.00026 | 1 |




Haplotypes representing the most frequent configurations found at the C4 *locus* in the HPRC dataset.

# Activities

https://hackmd.io/@AndreaGuarracino/H19Gn7VDc