Computational PANGenomics 2022

#CPANG22

Instituto Gulbenkian de Ciência, Portugal
Day 1 - 2022/05/23

Erik Garrison and Andrea Guarracino

Pangenome graph building

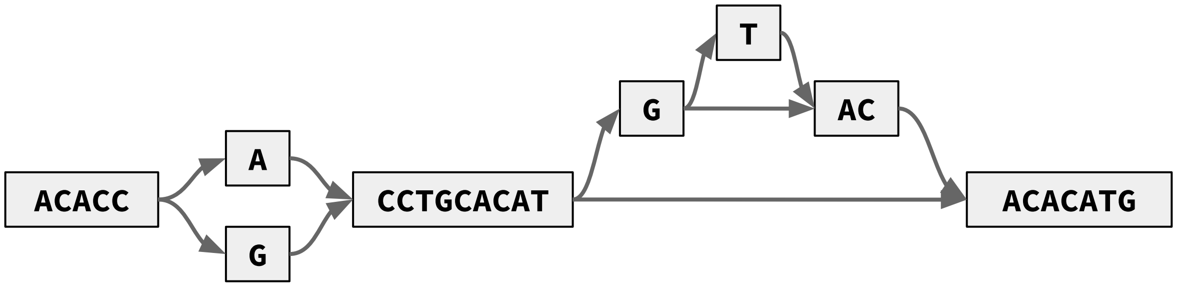
High-quality assemblies

```
ACACCACCTGCACATGACACACATG
ACACCACCTGCACATACACATG
ACACCACCTGCACATGACACACATG
ACACCACCTGCACATACACATG
ACACCACCTGCACATGACACACATG
ACACCGCCTGCACATGACACACATG
ACACCGCCTGCACATGTACACACATG
ACACCGCCTGCACATGACACACATG
```

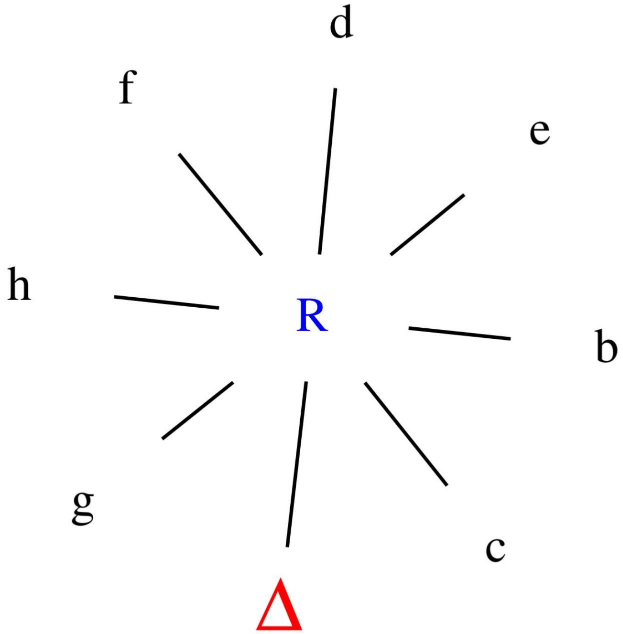
Alignment

```
ACACCACCTGCACAT GACACACATG
ACACCACCTGCACAT ---ACACATG
ACACCACCTGCACAT GACACACATG
ACACCACCTGCACAT ---ACACATG
ACACCACCTGCACAT GACACACATG
ACACCGCCTGCACAT GACACACATG
ACACCGCCTGCACAT GTACACACATG
ACACCGCCTGCACAT GACACACATG
```

Pangenome Graph



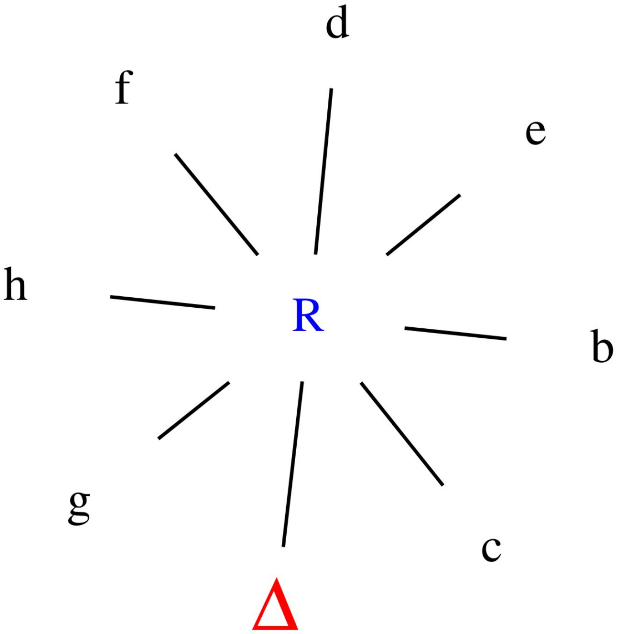
“Genomic” (reference-based) vs. Pangenomic models



aligning genomes to a reference

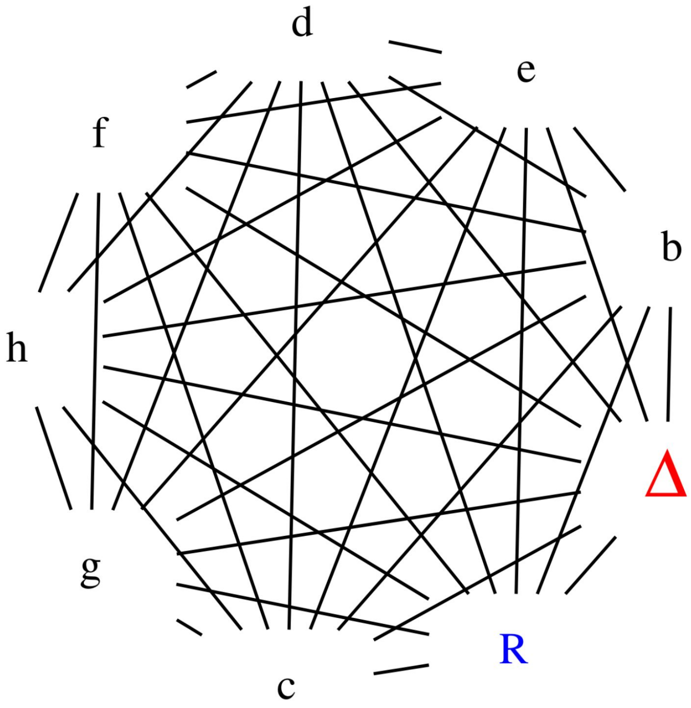
R = a reference
Δ = the “Nth” genome

“Genomic” (reference-based) vs. Pangenomic models



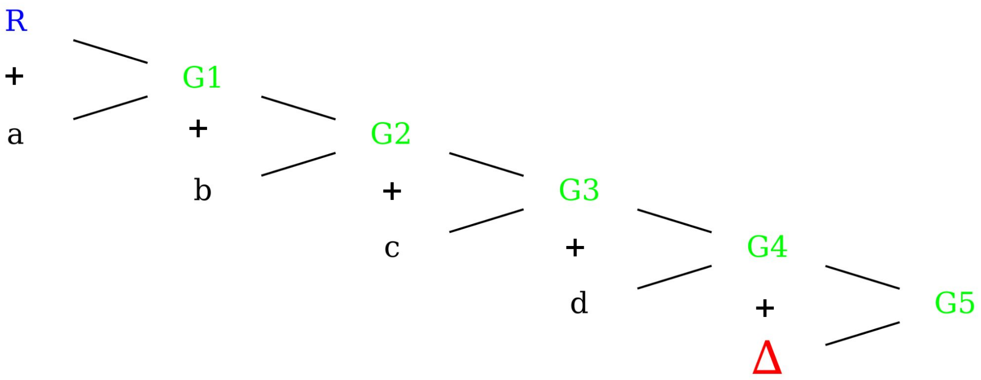
aligning genomes to a reference

R = a reference
Δ = the “Nth” genome

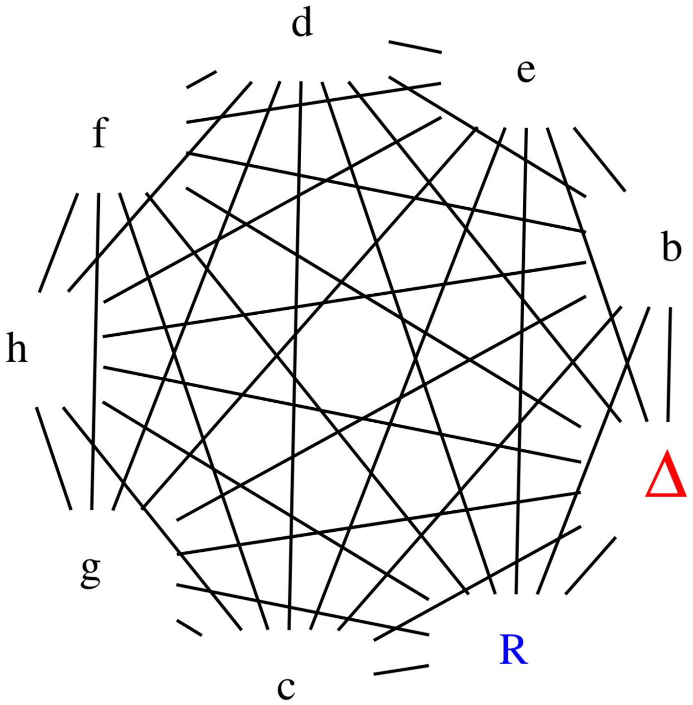


all-vs-all pangenome model

“Genomic” (reference-based) vs. Pangenomic models



progressive pangenome
model (minigraph)



R = a reference
Δ = the “Nth” genome

all-vs-all pangenome model

the PanGenome Graph Builder (PGGB)

https://github.com/pangenome/pggb

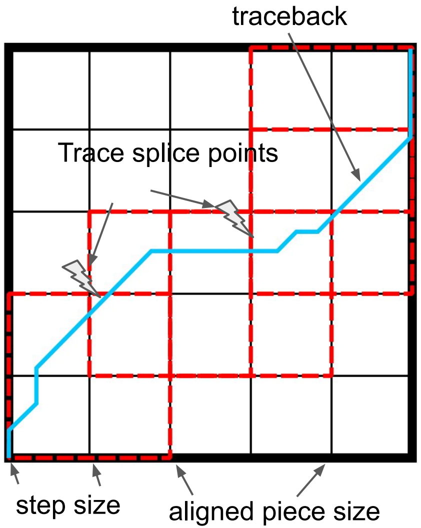
Solving the whole genome alignment problem in 3 steps.

the PanGenome Graph Builder (PGGB)

https://github.com/pangenome/pggb

Solving the whole genome alignment problem in 3 steps.

1) all-to-all alignment with **wfmash**



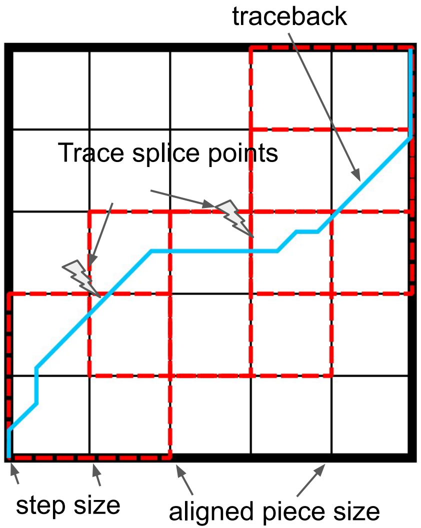
https://github.com/waveygang/wfmash

the PanGenome Graph Builder (PGGB)

https://github.com/pangenome/pggb

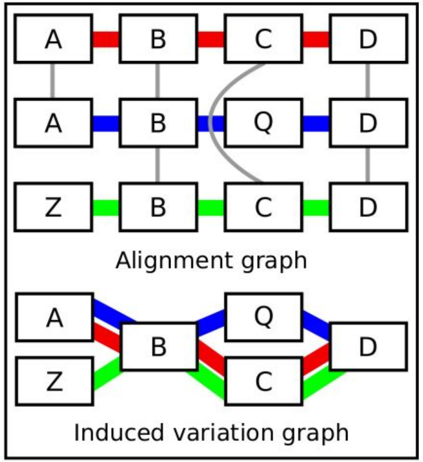
Solving the whole genome alignment problem in 3 steps.

1) all-to-all alignment with **wfmash**



https://github.com/waveygang/wfmash

2) graph induction with **seqwish**



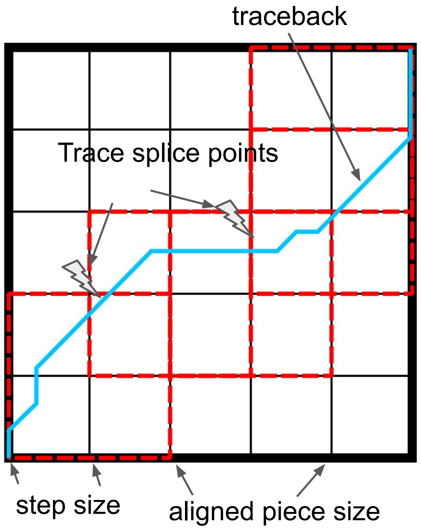
https://github.com/ekg/seqwish

the PanGenome Graph Builder (PGGB)

https://github.com/pangenome/pggb

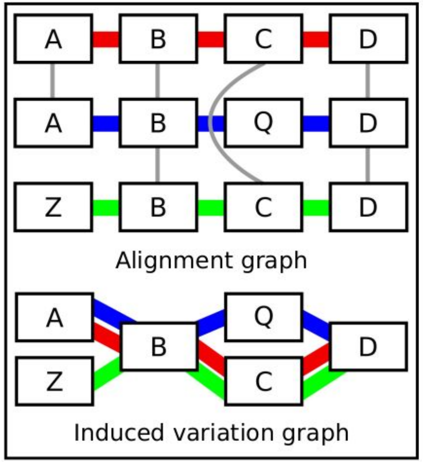
Solving the whole genome alignment problem in 3 steps.

1) all-to-all alignment with **wfmash**



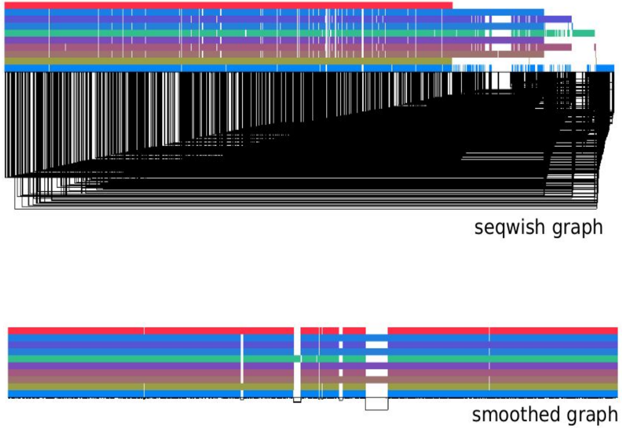
https://github.com/waveygang/wfmash

2) graph induction with **seqwish**



https://github.com/ekg/seqwish

3) normalization with **smoothxg**

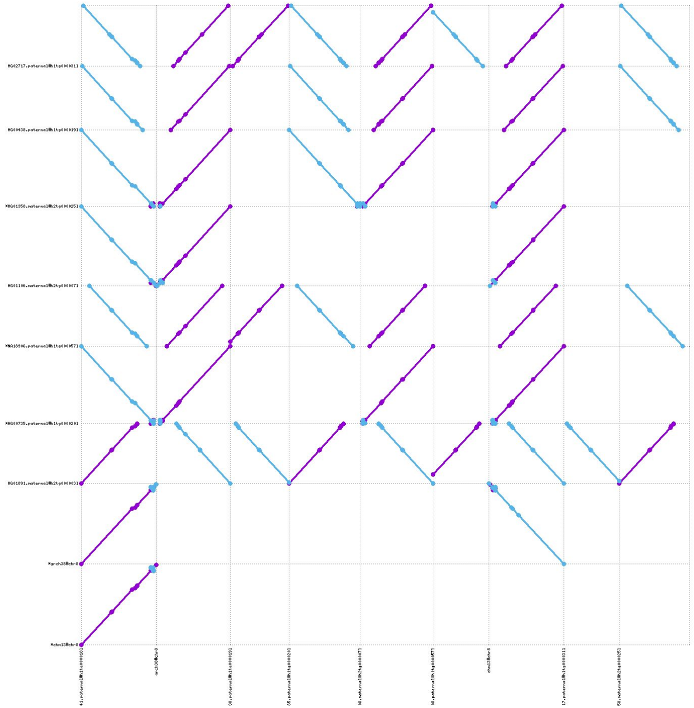


https://github.com/pangenome/smoothxg

All-to-all alignment with wfmash

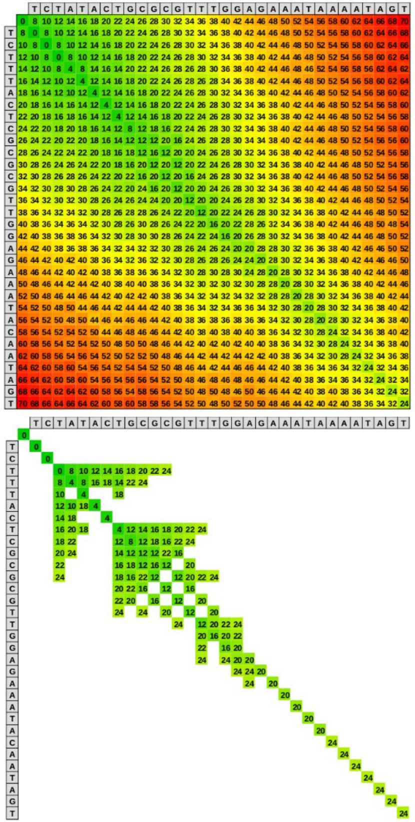
We first apply a heuristic homology mapping step (Jain et al., 2018) that efficiently finds regions of query and target sequences that are likely to be good alignments.

However, it has no facility to derive the precise base-level alignments.



Dot plot representing the pairwise mappings of human centromere-spanning contigs of chromosome 8.

All-to-all alignment with wfmash

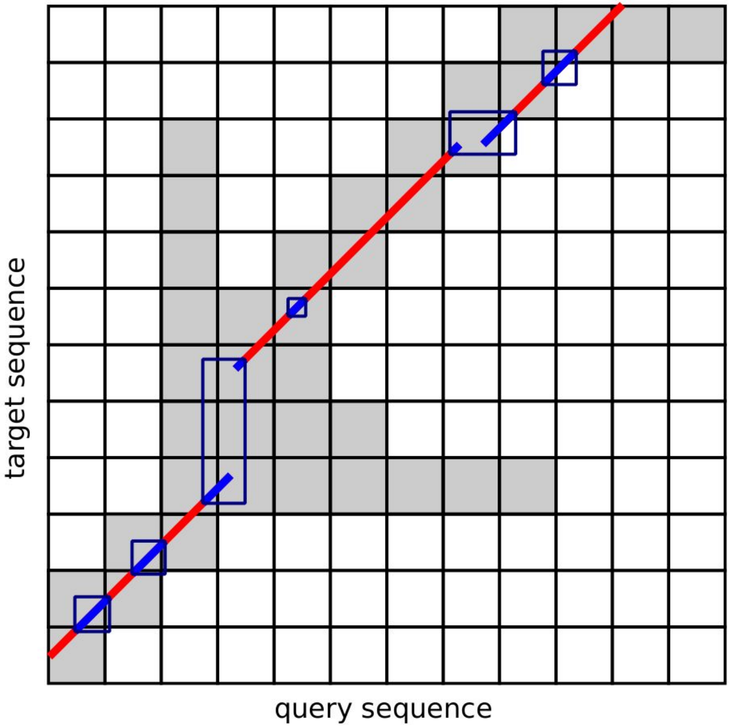

TCTATACTGCGCGTTTGGAGAAATAAAATAGT
TCTATACTGCGCGTTTGGAGAAATAAAATAGT

Classical pairwise alignment (Needleman-Wunsch, Smith Waterman).

Wavefront Alignment (WFA) (Marco-Sola et al., 2020)

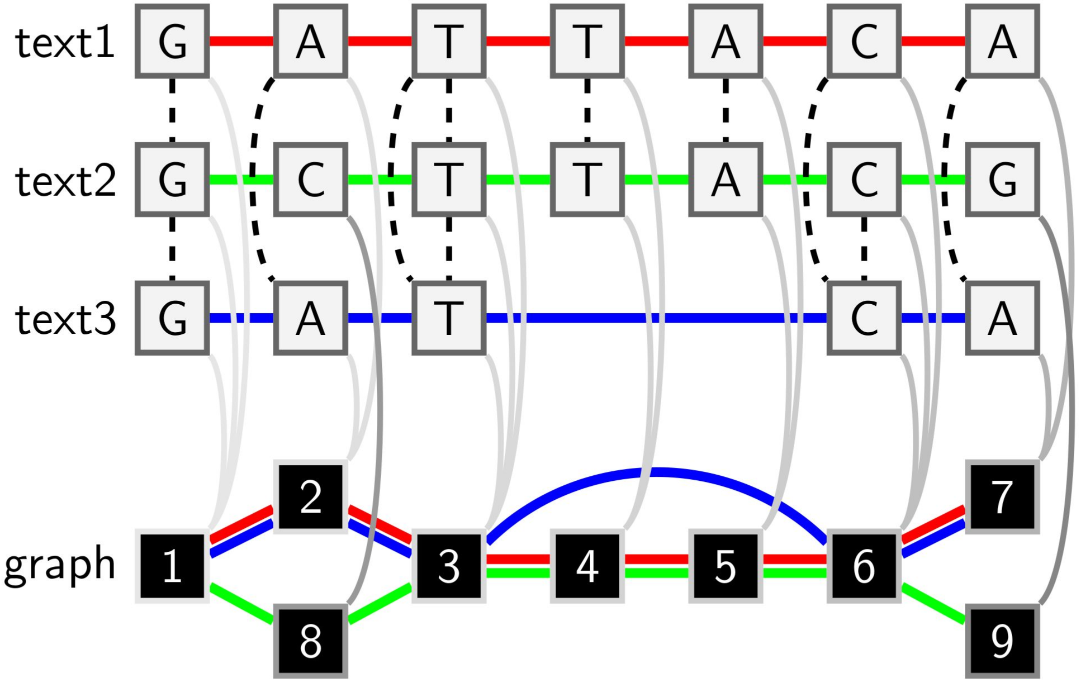
https://github.com/smarco/WFA2-lib

wflign



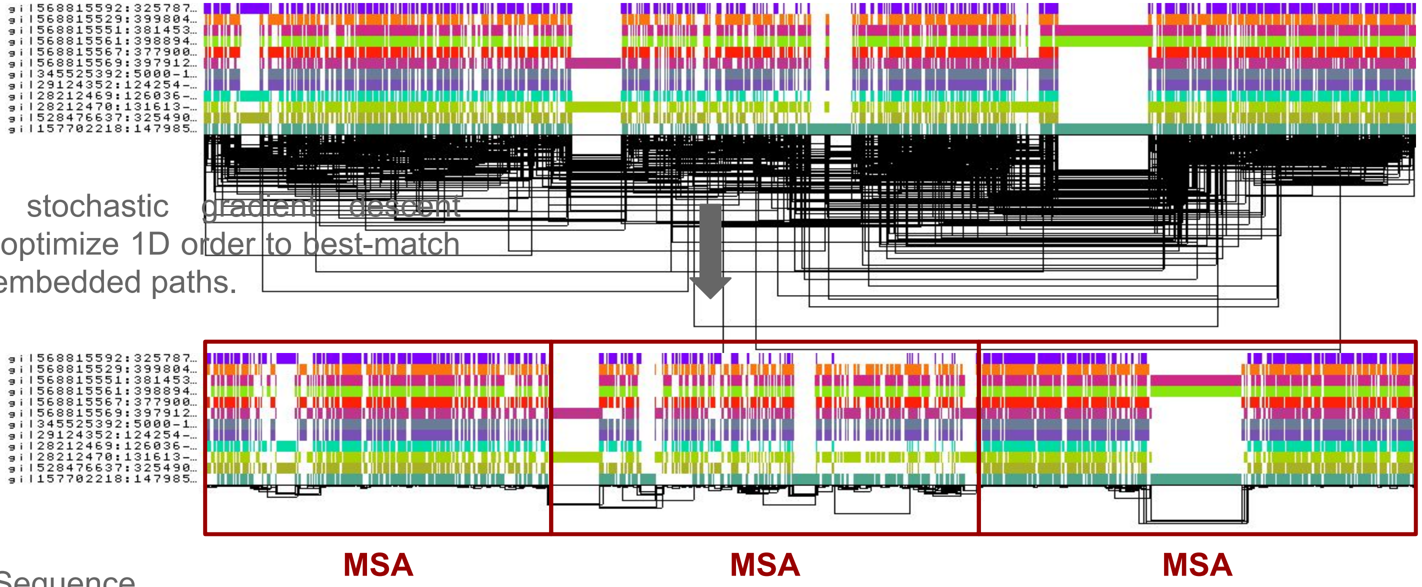
https://github.com/ekg/wflign

Graph induction with seqwish



Normalization with smoothxg

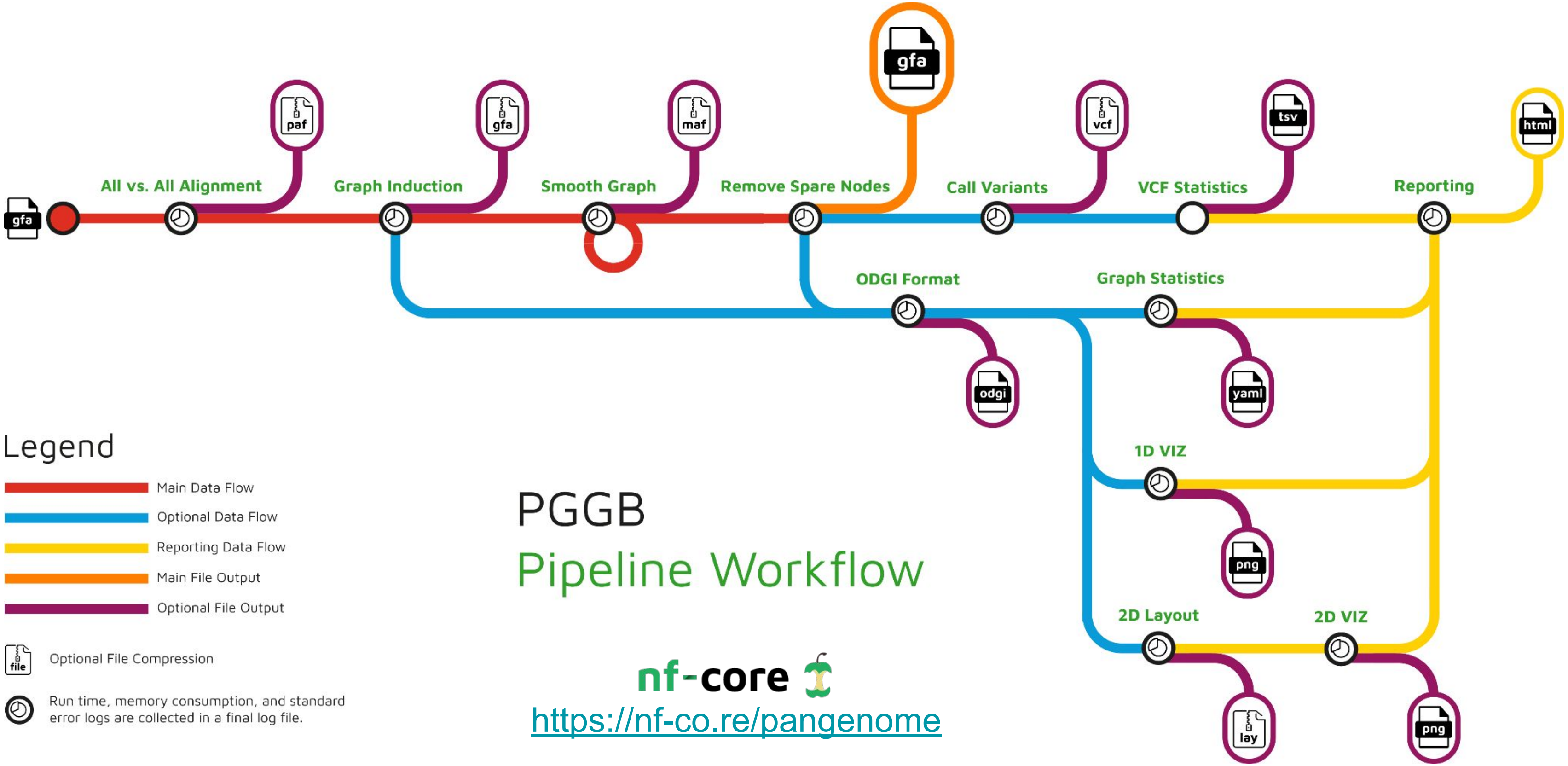
Pangenome graph with 12 ALT sequences of the HLA-DRB1 gene from the GRCh38 reference genome.



Path-guided stochastic gradient descent algorithm to optimize 1D order to best-match positions in embedded paths.

Multiple Sequence Alignment (MSA) over the sorted graph, locally

Workflow



https://nf-co.re/pangenome

Main parameters

- **`-s/--segment-length`**, segment length for mapping,
- **`-p/--map-pct-id`**, percent identity for mapping/alignment,
- **`-n/--n-mappings`**, number of mappings to retain for each segment.

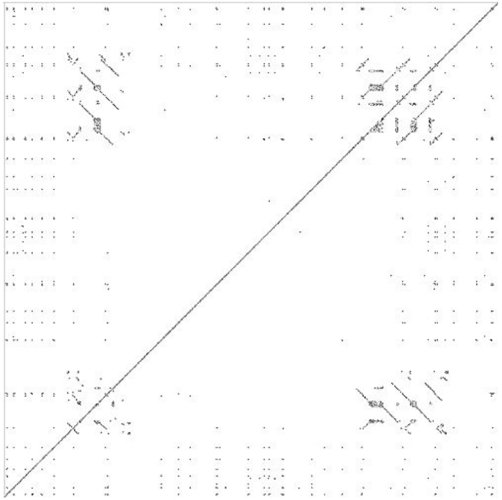
Activities

https://hackmd.io/@AndreaGuarracino/S1Qbe27v5

All-to-all alignment with wfmash

Current alignment methods do not scale to the much harder problem of mapping many reference-quality genomes to each other.

Tools based on seed-and-extend chaining of minimizers and k-mers must consider all candidate chains of a given length, affecting performance and downstream analyses.



Dot plot representing the alignment of an *in silico* mutated beta-defensin locus (divergence 1%) against the CHM13 reference. Alignment performed with minimap2. Image produced with pafplot.