



HAL
open science

Bioinfomics days : introduction to variation graphs

Benjamin Linard

► **To cite this version:**

| Benjamin Linard. Bioinfomics days : introduction to variation graphs. 2024. <hal-04828607>

HAL Id: hal-04828607

<https://hal.inrae.fr/hal-04828607v1>

Preprint submitted on 10 Dec 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

➤ Introduction to variation graphs (VG)

Benjamin Linard

benjamin.linard@inrae.fr
miat.inrae.fr/teams/saab

05-12-2024



RÉPUBLIQUE
FRANÇAISE

*Liberté
Égalité
Fraternité*

INRAE

MIA
TOULOUSE

I. Context



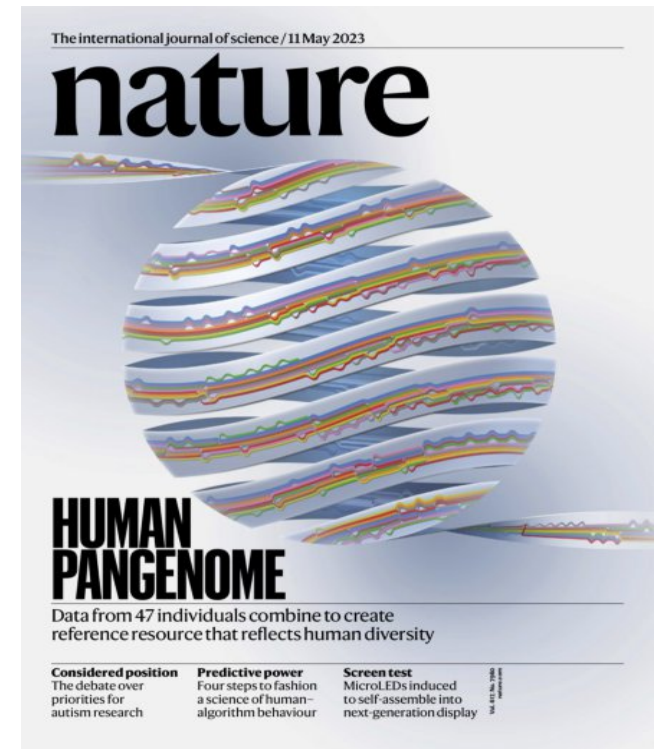
2001

1st « complete »
genome



2010

First overview of
genome variations
diversity



2023

Paradigm shift
Analysis contextualised
with full known
genome diversity

I. Context

Why now ?

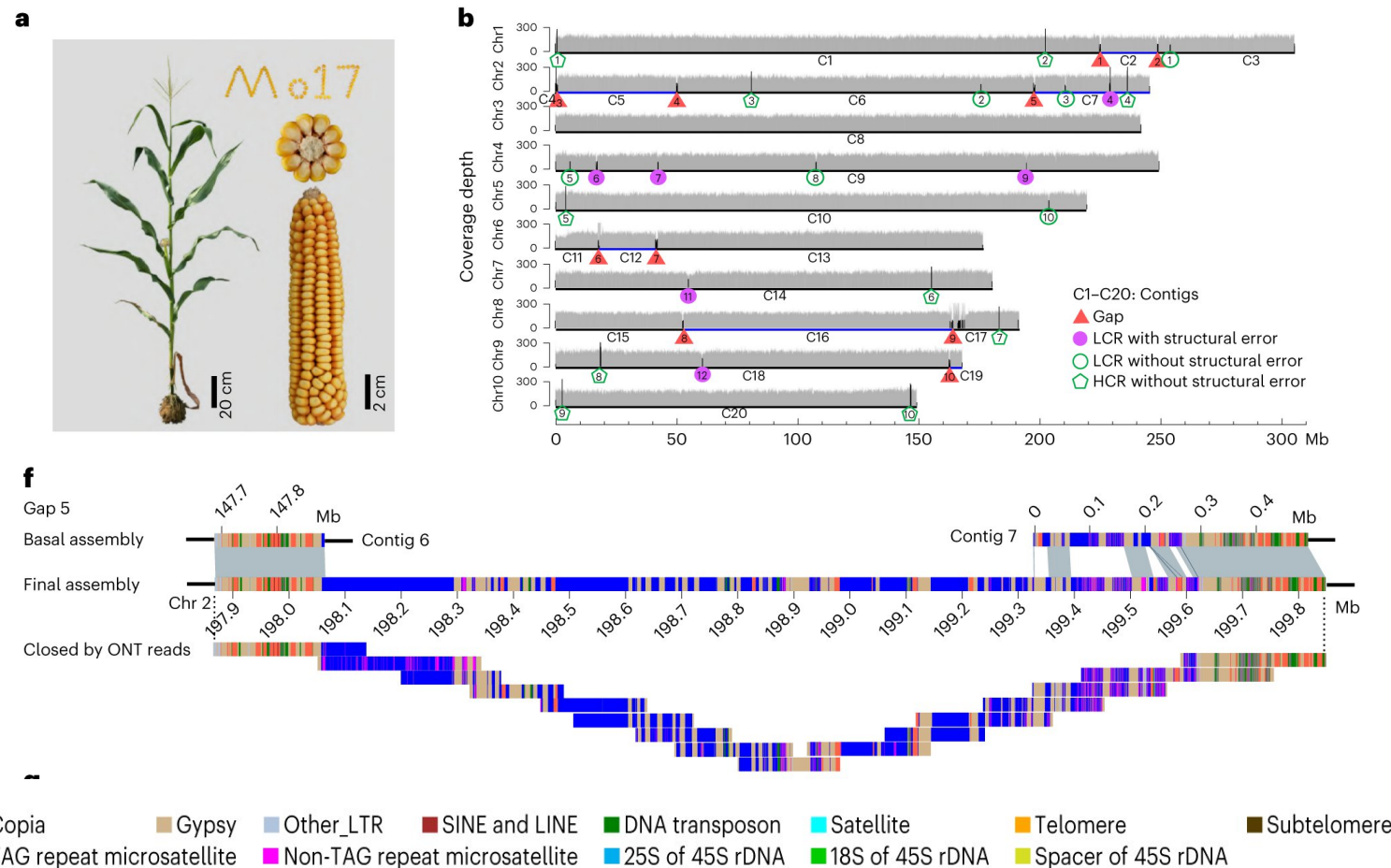
Recent advances
in genome sequencing.

- Many long read technologies (Hi-Fi, ONT, HiFi, 10x)
- Hi-C : 3D data improves long-distance scaffolding
- Efficient hybrid assemblers: Hifiasm, CANU, 3DDNA ...

**Haplotype-resolved,
telomere to telomere,
full genome assemblies !**

Chen et al, June 2023.

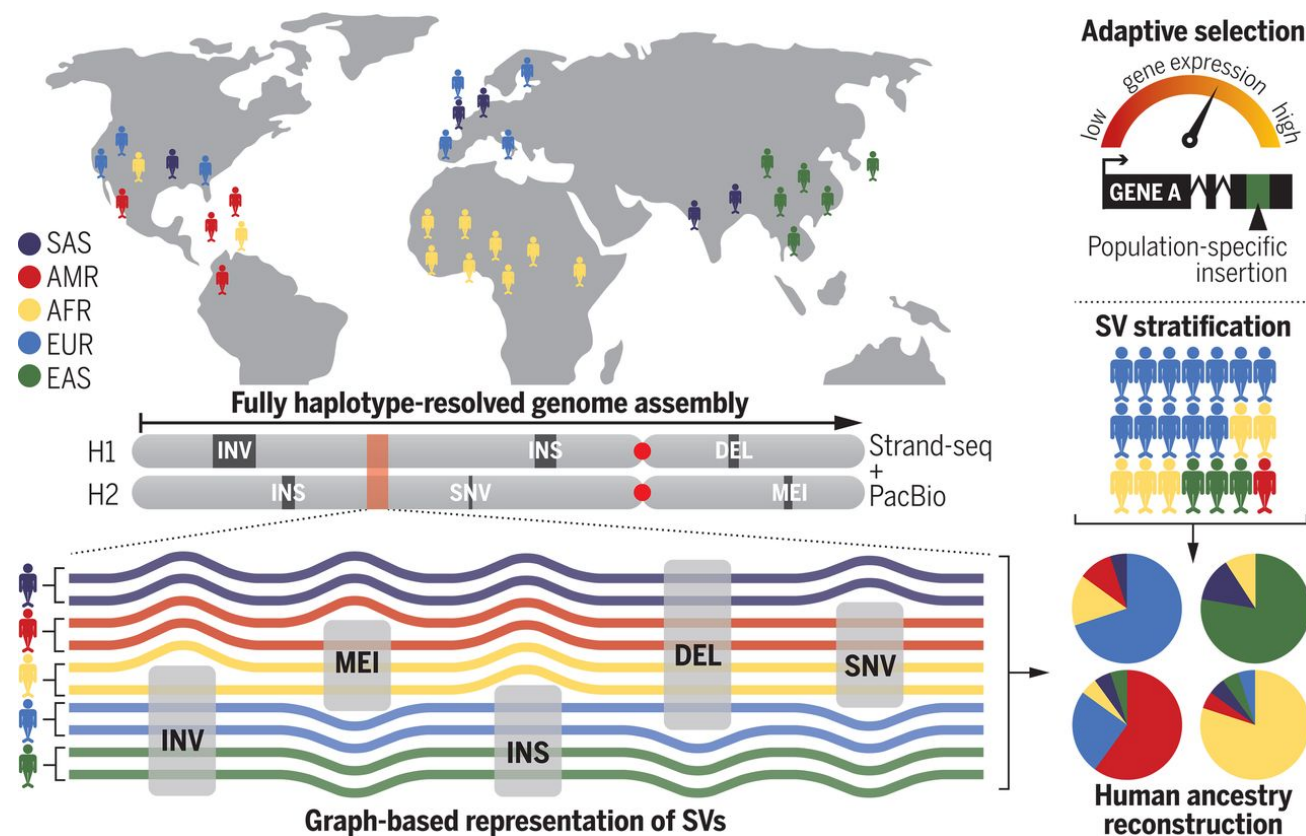
“The 2,178.6 Mb T2T Mo17 genome with a base accuracy of over 99.99% unveiled the structural features of all repetitive regions of the genome.”



I. Context

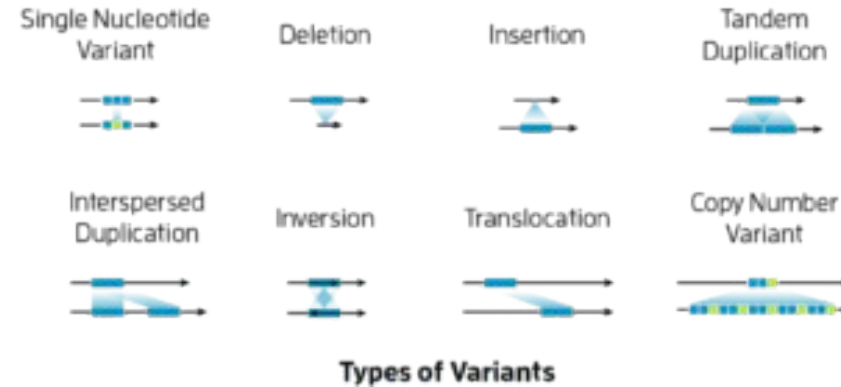
Paradigm change : genomic analyses can be contextualised with all known genome diversity

- “The current version of the reference genome (GRCh38) is estimated to miss up to 10% of our species genetic information” (SS Sherman RM, 2020)



I. Context : structural variations (SV)

- A major step:
better access to genome
Structural variants (SV)



- Unexplored impact of SVs

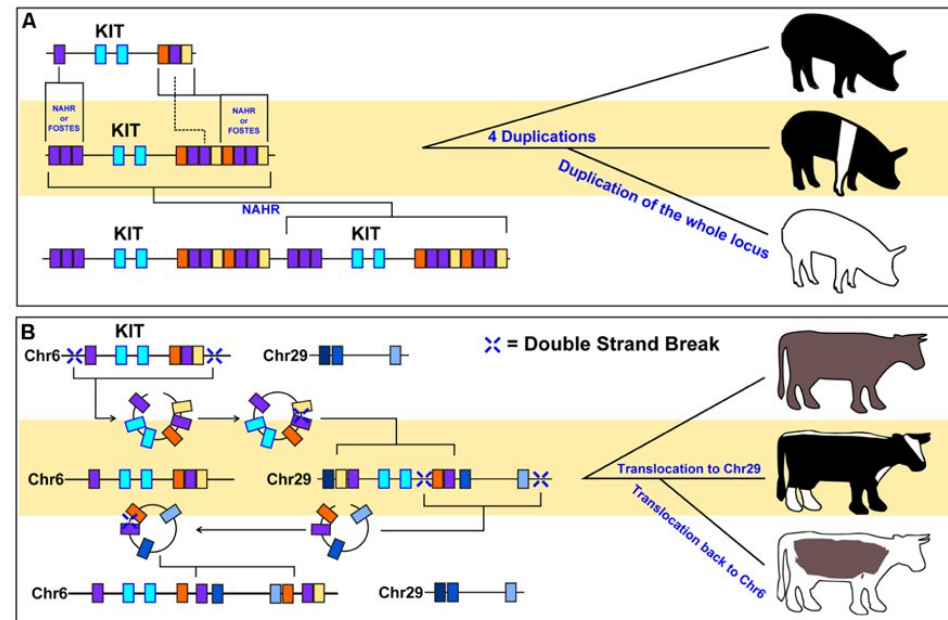
Dosage effects



**Disruptions:
Gene or
regulation**

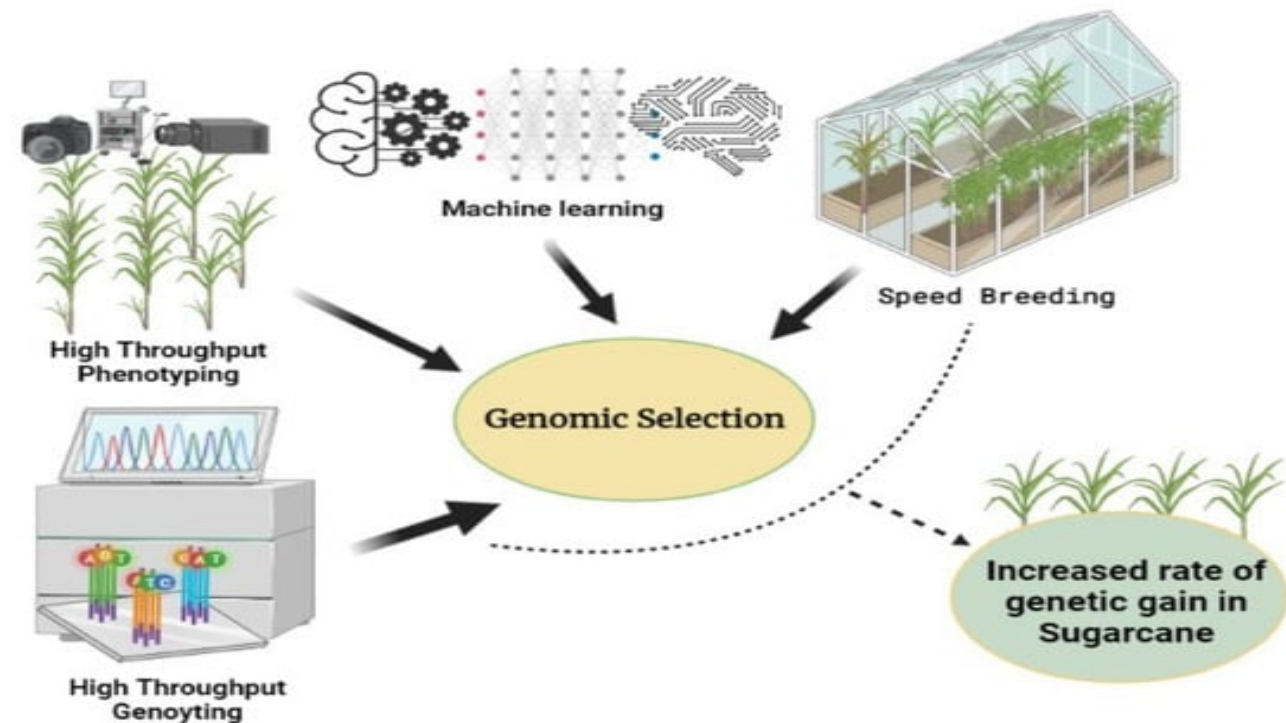


Positional effects



I. Context : agronomy and selection

- **New context for data generation**
 - HiFi & Hi-C technologies
 - « telomere to telomere » genomes
 - Assembly is becoming a routine step
- **New opportunities for genetics**



I. Context : agronomy and selection

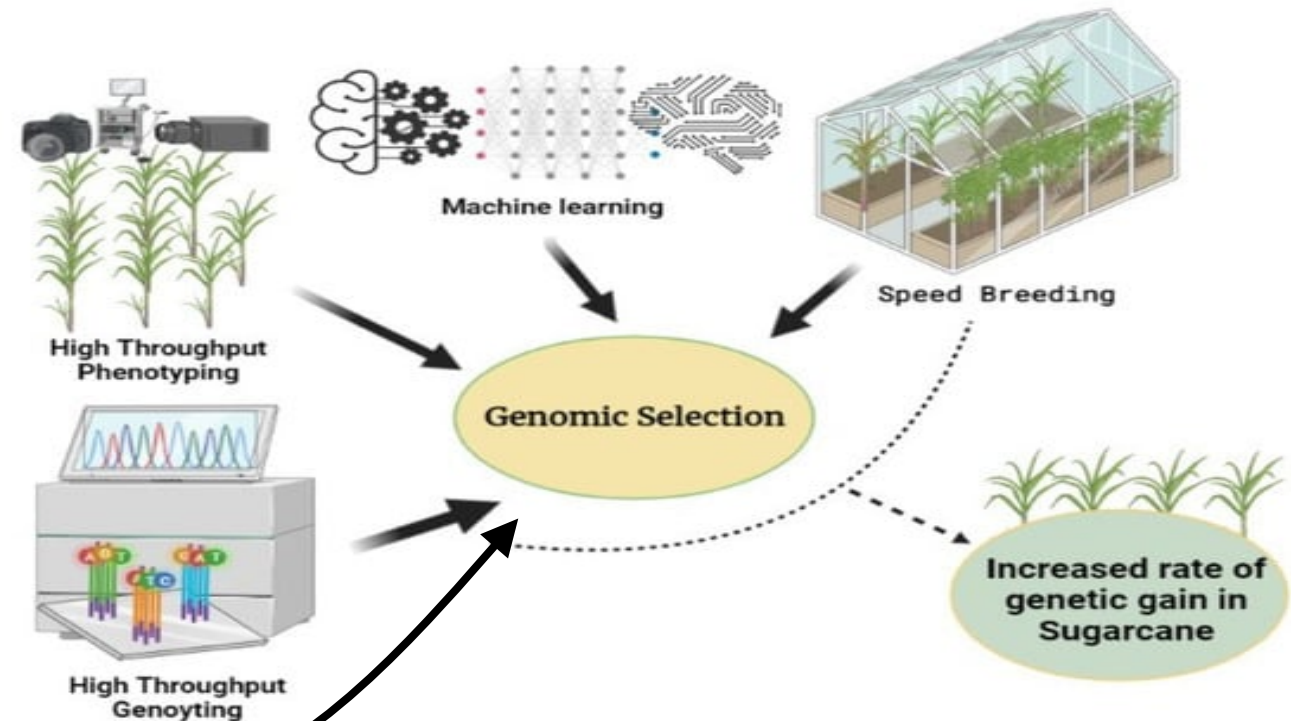
- **New context for data generation**

- HiFi & Hi-C technologies
- « telomere to telomere » genomes
- Assembly is becoming a routine step

- **New opportunities for genetics**

- Complete genomes
=> SV extraction and analysis
- Many individuals
=> variant frequencies and histories

**Genomic diversity
of the species**



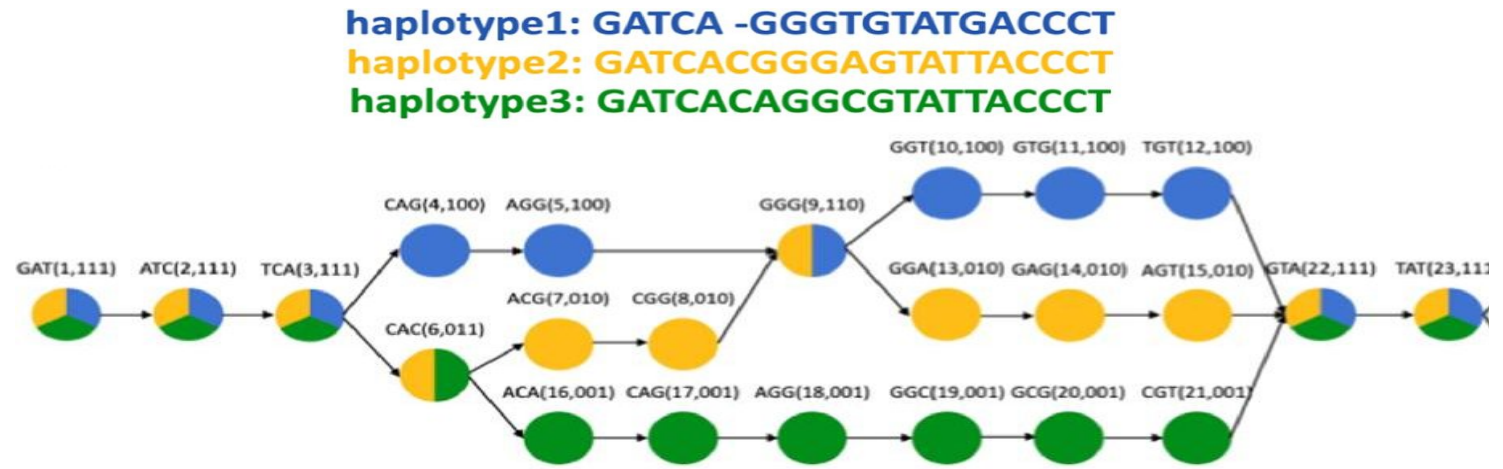
II. Models

- Modelling genome diversity with a graph

De Bruijn graphs
(compacted/ colored)

Microbiomes
(mainly)

(>1000 small genomes)

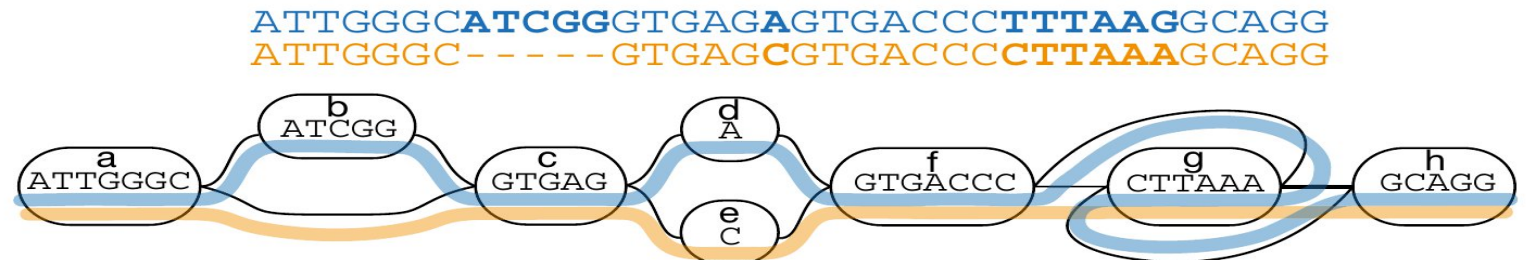


Guo, J. *et al.* (2021)

Variation graphs

Eukaryote genomes
(mainly)

(10 to 100 large genomes)



II. VG model : avoiding reference biases

(a)

Ref. `ACGGTTAAGGGCGATCG--CTCGTTTT`
`ACGGTTAAG--CGATCG--CTCGTTTT`
`ACCGTTAA----GATCGAACTCG----`
`ACCGTTAAGGGCGATCGAA----TTTT`

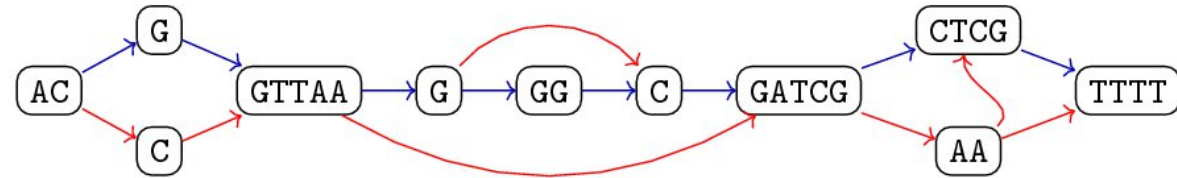
Reads:

`ACCGTTAAGCGA`
`TCGAATTTT`

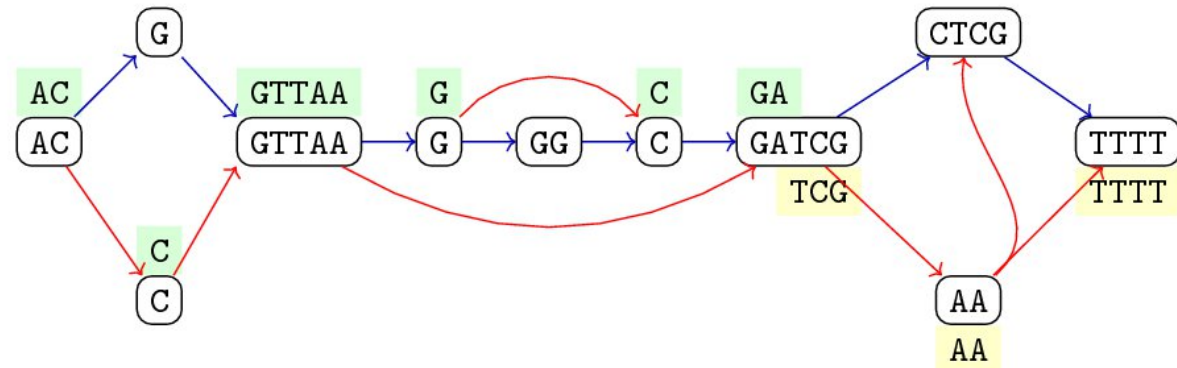
(c)

`ACCGTTAAGCGA`
`ACCGTTAAGGGCGATCGCTCGTTTT`
`TCGAA--TTTT`

(b)

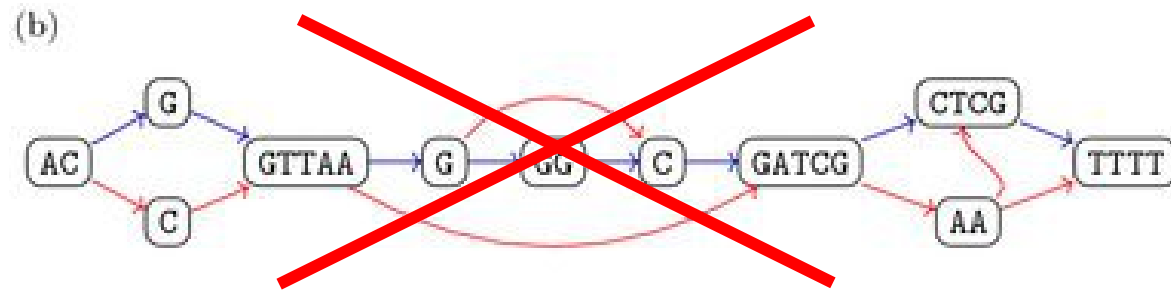


(d)



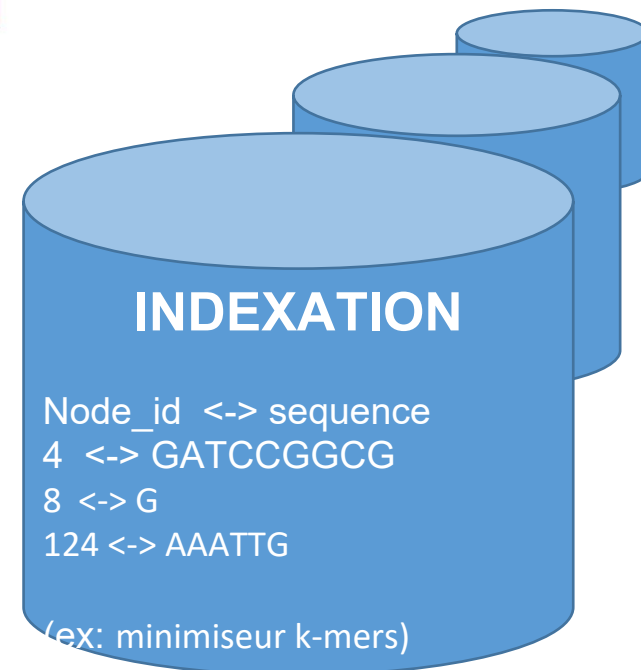
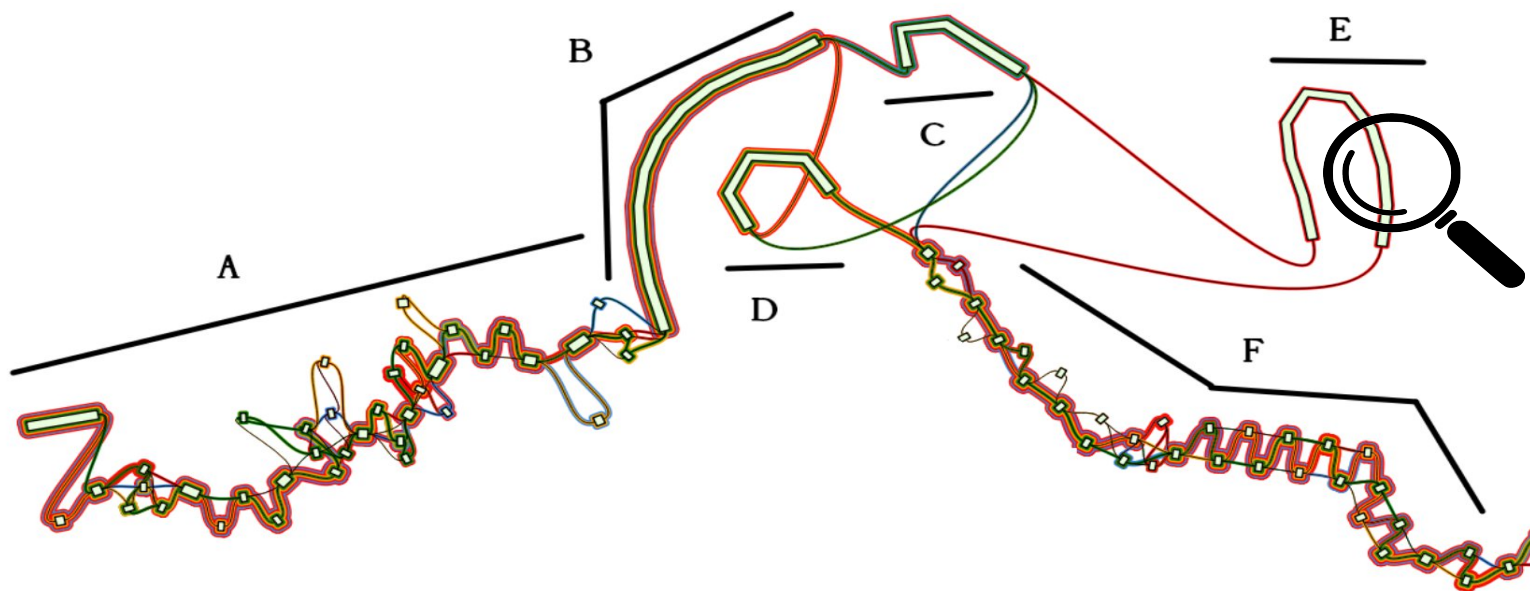
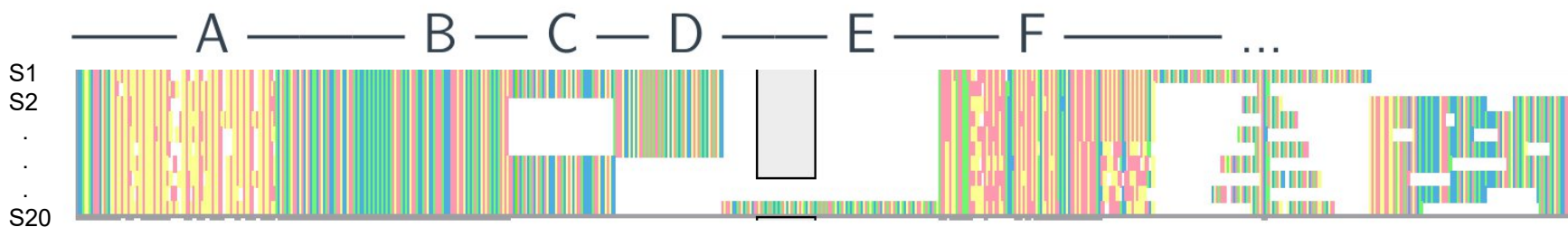
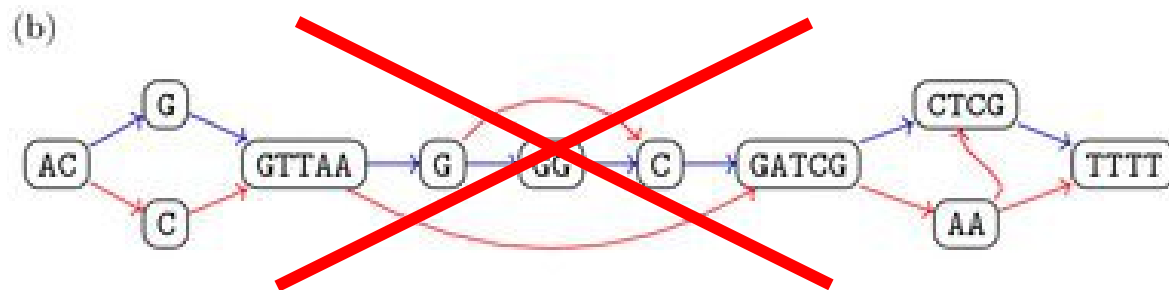
II. VG model

- In practice ...



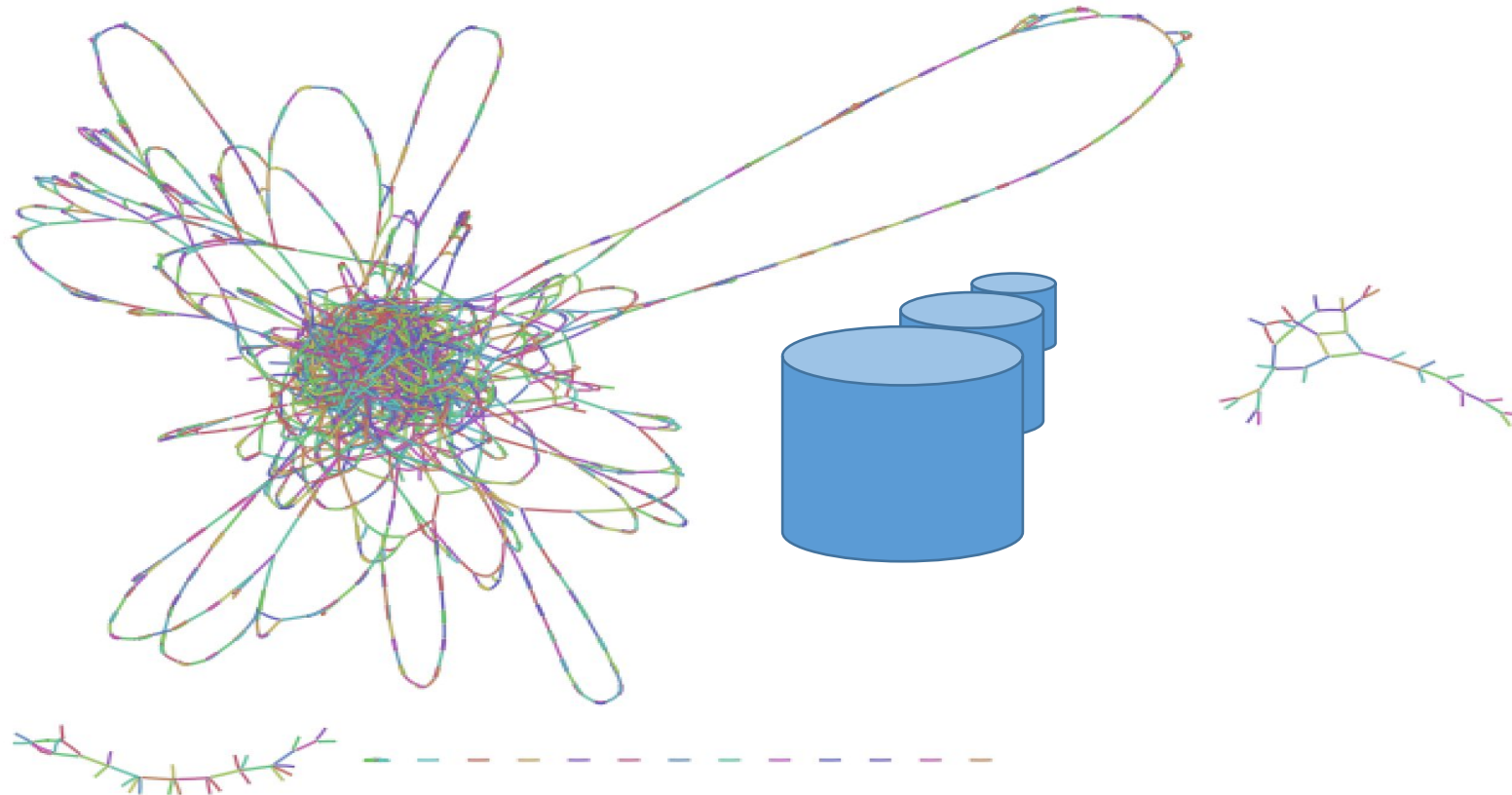
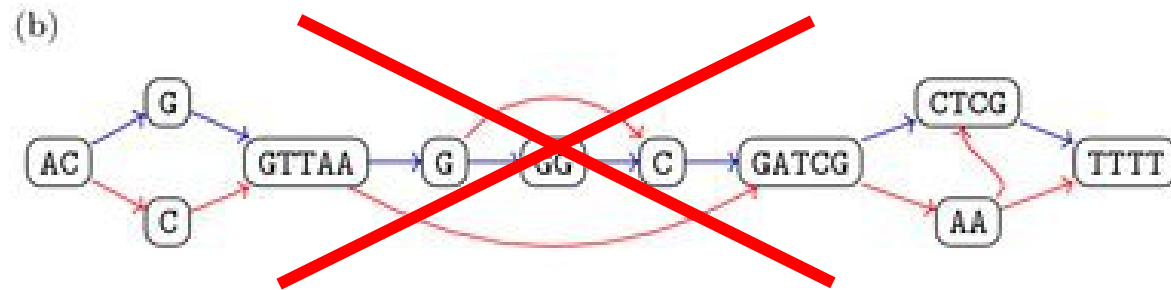
II. VG model

- In practice ...



II. VG model

- In practice ...



II. VG models : challenges

- **Challenges related to the model and its exploitation**

- Graph dimensions :**

- 12 bovines assemblies, $>10^7$ nodes (Leonard AS et al, 2023)
 - problems related to memory / indexation
 - nœud : associé de 1 à n kilobases

- Topology :**

- potentially nodes with high degree
 - complex bubbles in some regions
 - cycles (many algorithms are not polynomial anymore)

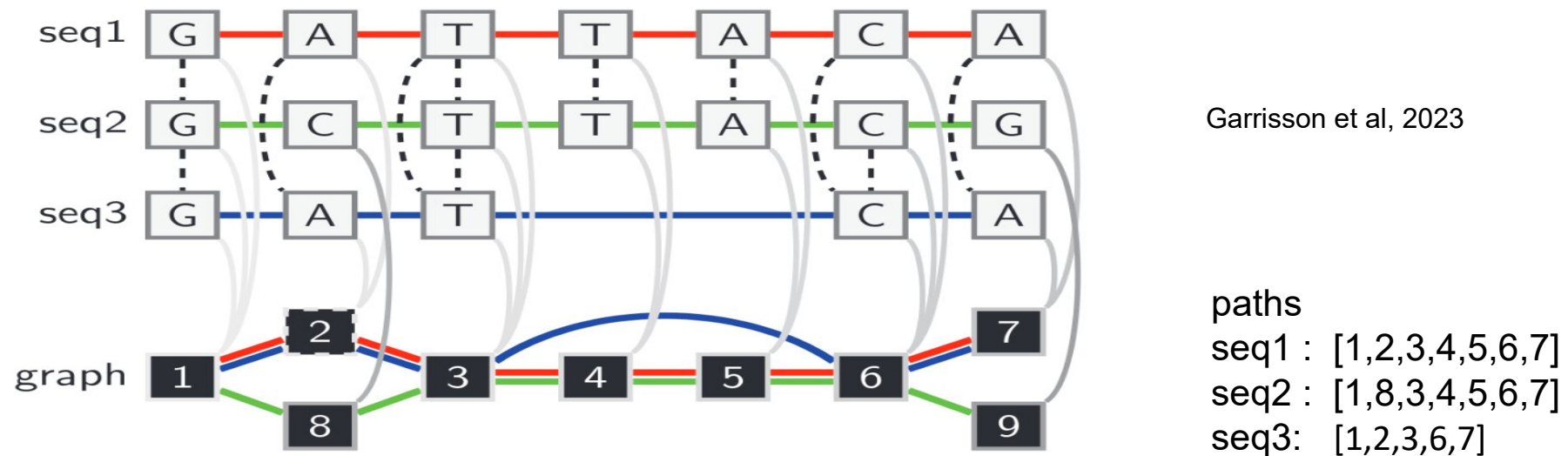
- Applications :**

- Up to which level of genetic diversity : how many haplotypes ? beyond the species level ?
 - Adapting bioinformtic tools to this new representation

III. Construction : global approach

- **Construction of a variation graph**

- Inputs :
 - N a set of high-quality assemblies, pairwise aligned
 - A the set of per-character pairwise homology (alignments)
- Outputs : An oriented graph (string graph) $G = \{V, E, P\}$
where P , is the set of path representing the integrated assemblies
- Aim : G includes N and every pairwise relationships described by the genome alignments
(G cannot contain more information than in A)



III. Construction : available tools

In its own category:

- ▶ **Variation Graph (VG)** (Garrison et al, 2018)

Genome alignment, graph construction, Post-processes.

- ▶ **Minigraph (MG)** (Li et al, 2020)
- ▶ **Minigraph-Cactus (MGC)** (Hickey et al, 2023)
- ▶ **PanGenome Graph Builder (PGGB)** (Garrison et al, 2023)

III. Construction : VG

▶ INPUT:

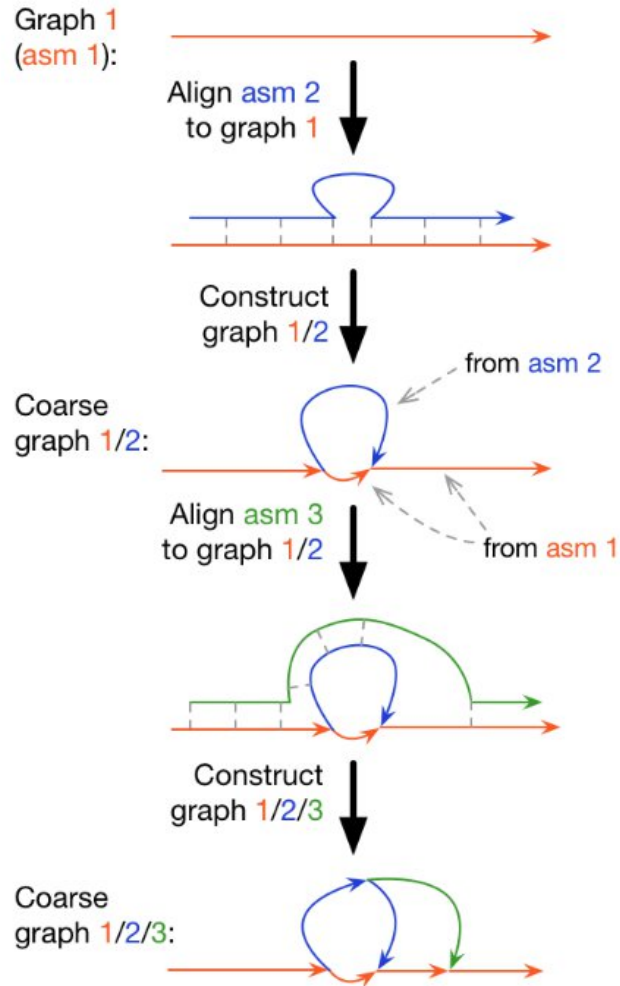
- ▶ A complete reference genome (fasta)
- ▶ Sets of variants, as VCF files relative to this reference.
- ▶ Or multiple sequence alignments (fasta, clustal)

▶ INCREMENTAL CONSTRUCTION:

- ▶ Each variant described in the VCF adds a topological change in the graph.
- ▶ Every variant is relative to the reference.
- ▶ Excepted reference, assemblies not compulsory (ex: BED>VCFs from chip-seq)

▶ TOY EXAMPLE: gtpb.github.io/CPANG18/pages/toy_examples

III. Construction : Minigraph



▶ INPUT:

- ▶ A complete reference genome (fasta)
- ▶ A set of other haplotypes (fasta)

▶ INCREMENTAL CONSTRUCTION:

- ▶ Haplo 2 mapped to ref: output $\rightarrow G_{[ref,2]}$
- ▶ Haplo 3 mapped to $G_{[ref,2]}$: output $\rightarrow G_{[ref,2,3]}$
- ▶ ... etc ...

III. Construction : Minigraph-cactus

▶ INPUT:

- ▶ At least 1, complete, high-quality reference genome (fasta)
- ▶ A set of assemblies for other individuals (fasta)
- ▶ (Optional) A tree describing the relationship between individuals

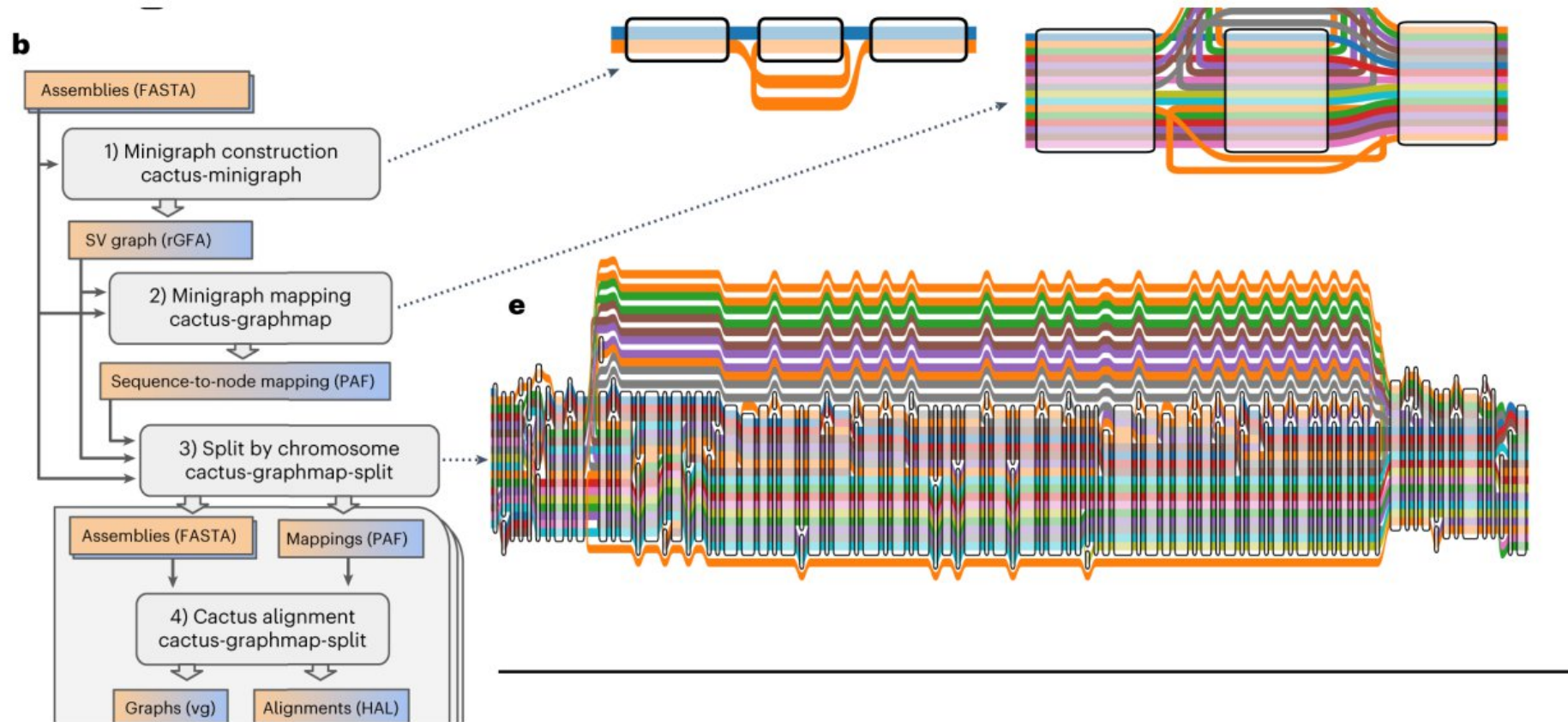
▶ INCREMENTAL CONSTRUCTION:

1. Minigraph launched to get high-level SV graph
2. Assemblies re-aligned to this backbone with Progressive-Cactus → adds nodes up to SNP level
3. Post-processes to simplify/modify the graph.

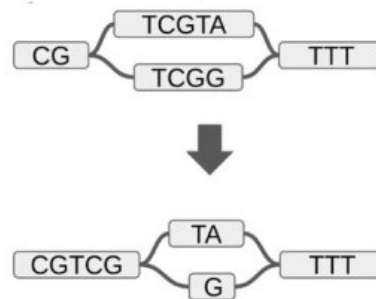
▶ WARNING:

- ▶ Minigraph min SV length is FIXED (programmatically, 50bp, 09/2023)

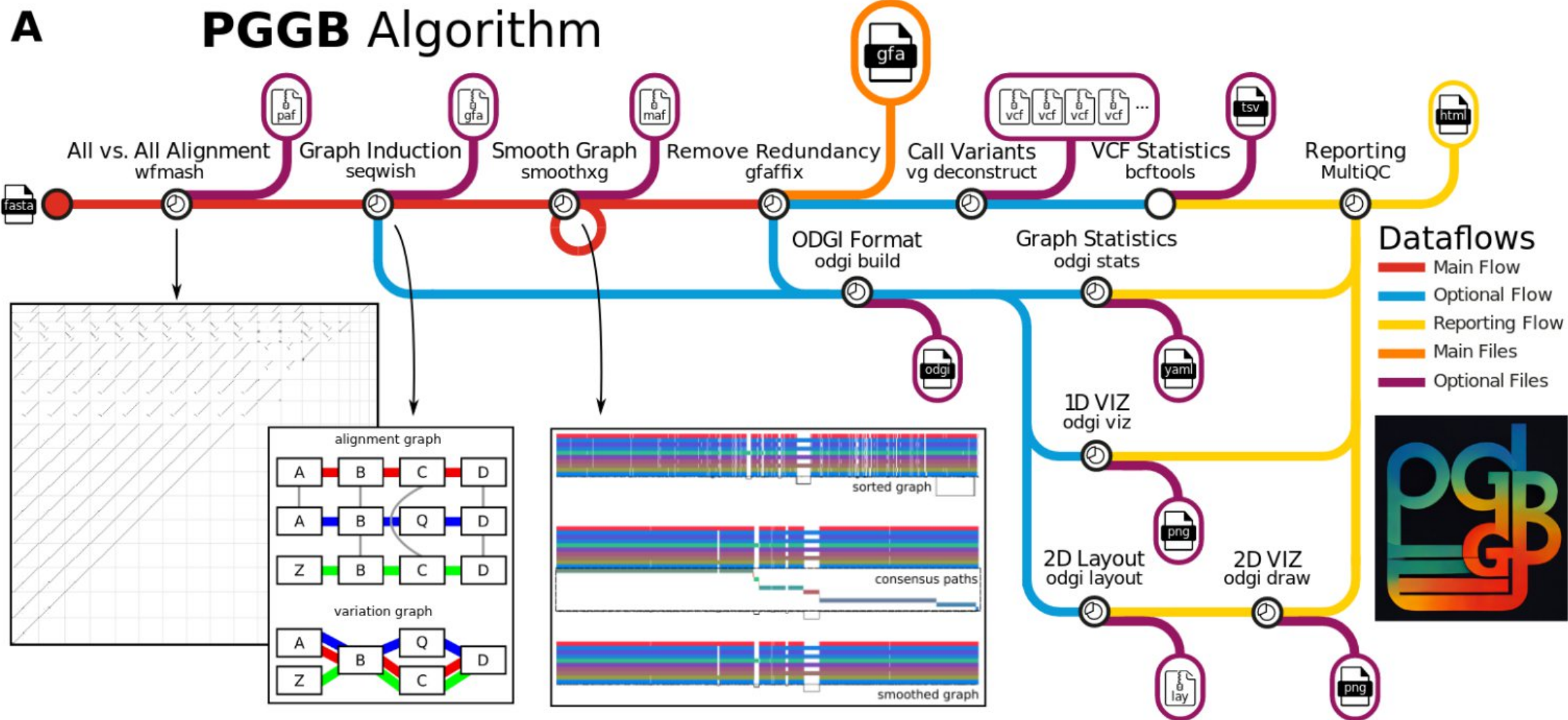
III. Construction : Minigraph-cactus



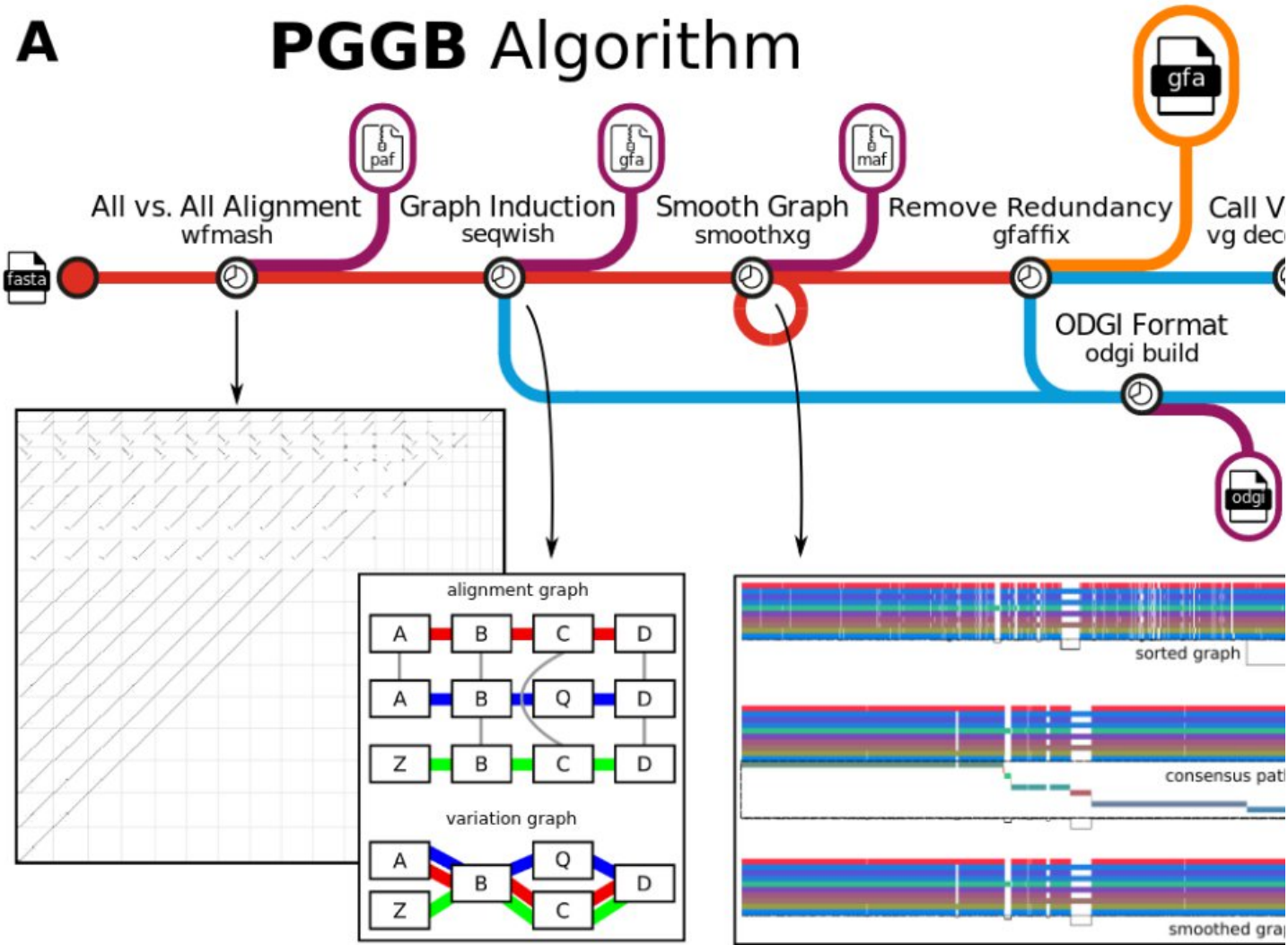
Post-process example: GFAffix tool



III. Construction : PGGB



III. Construction : PGGB



III. Construction : PGGB

▶ INPUT:

- ▶ A set of assemblies, one per chromosome.

▶ PROCESSUS:

1. Pairwise alignments for all assemblies with wfmash (Guarracino et al, 2023)
2. Graph induction with Seqwish (Garrisson et al, 2023)
3. Graph "smoothing" with smoothxg (not published, not documented)
4. Post-processes with GFAffix, ODGI ...

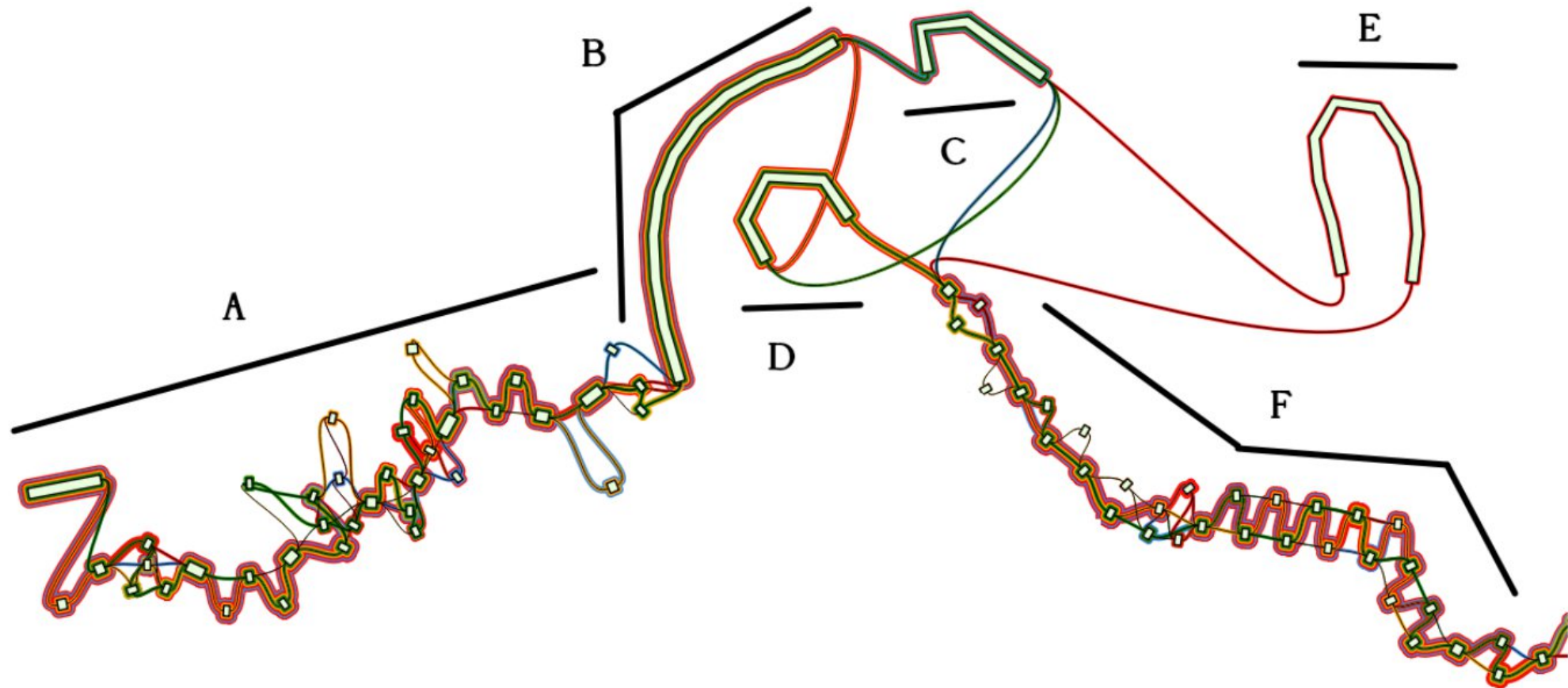
▶ COMMENTS:

- ▶ Similar to MGC, PGGB sets lots of default parameters for you.
- ▶ But logs are more clear, launched subprocesses are logged ...
- ▶ Each step can be run independantly and then parameters changed manually.
- ▶ Documentation / tutorials are not always clear.
- ▶ Smoothxg is not published, but does MANY things.

III. Construction : PGGB

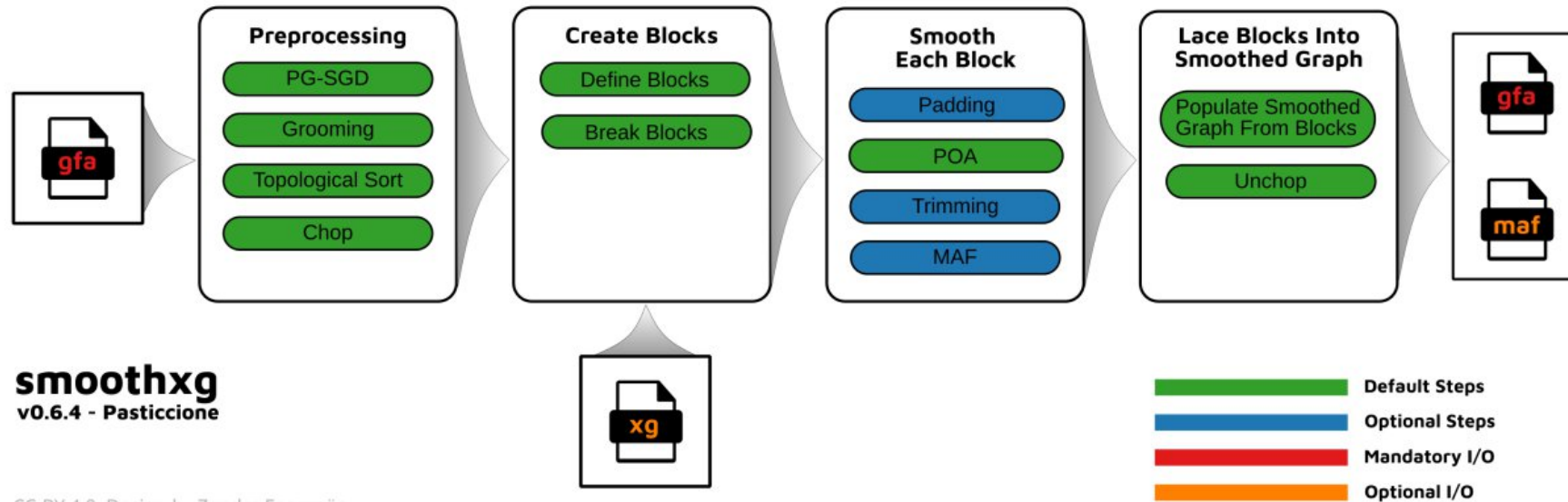
➤ Seqwish output IS the expected graph

But this is NOT
the final PGGB graph.



III. Construction : PGGB

Smoothxg role described recently in Garrisson et al, 2024. Nat Methods.

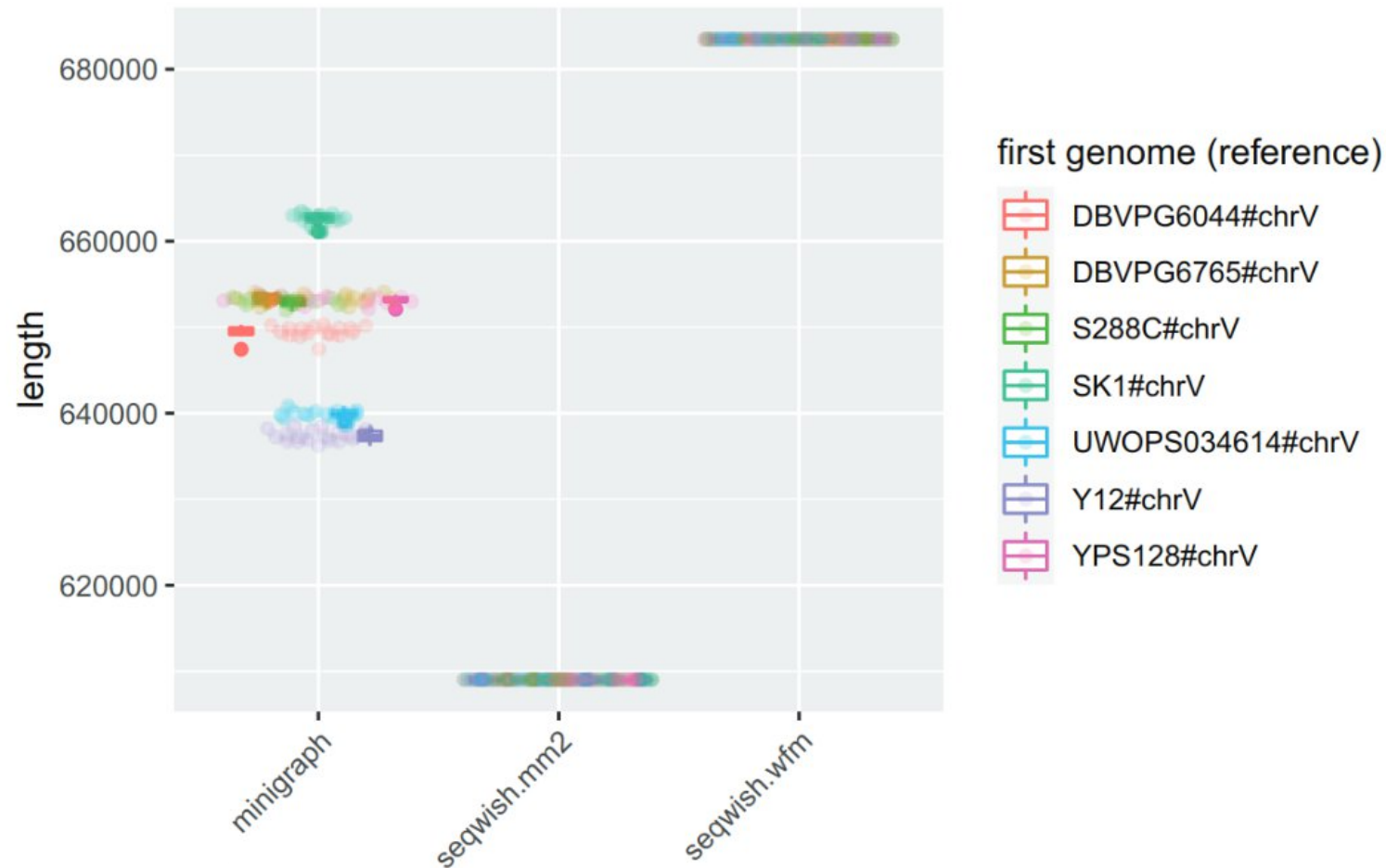


Note: Graph sorting then POA, which is similar to minigraph-cactus.

III. Construction : fundamental difference

Impact of genome order

Reminder: The tree generated (or given) to Minigraph & Minigraph-Cactus guides the iterative assembly alignment. (Figure from seqwish paper)



III. Construction

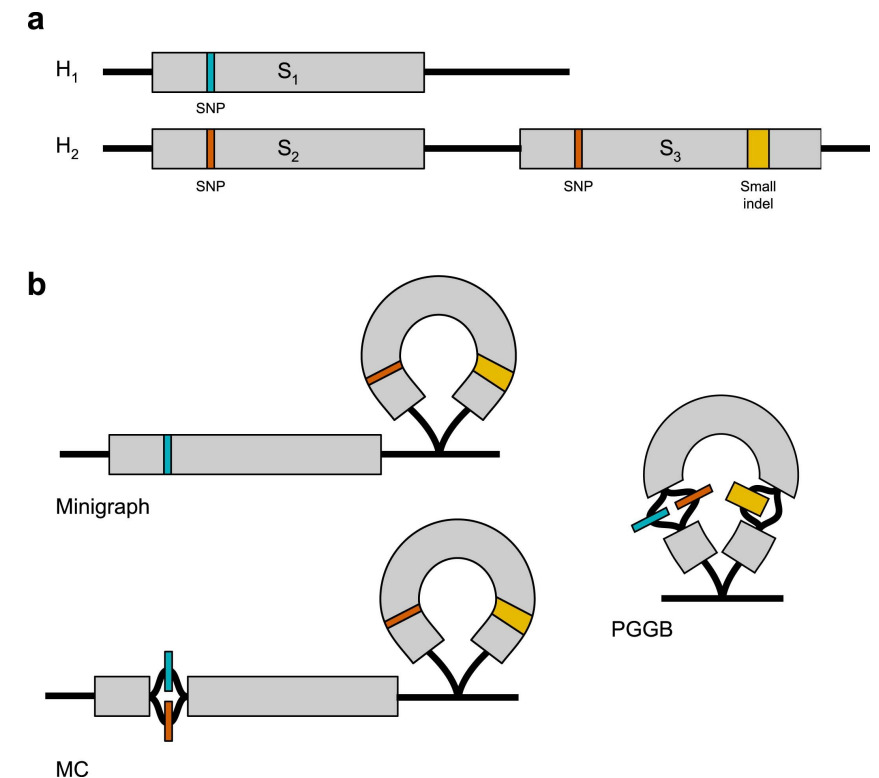
► Andreace et al 2023, <https://doi.org/10.1186/s13059-023-03098-2>

Metric	Bifrost	pggb	Minigraph-Cactus	Minigraph	mdbg
1) Construction speed	● ● ○	● ○ ○	● ○ ○	● ● ○	● ● ●
2) Variations	● ● ●	● ● ●	● ● ●	● ● ○	● ● ○
3) Scalability	● ● ●	● ○ ○	● ○ ○	● ● ○	● ● ●
4) Editability	● ● ●	● ● ○	● ○ ○	● ● ○	● ○ ○
5) Stability	● ● ●	● ○ ○	● ○ ○	● ● ○	● ● ●
6) Accessibility by downstream applications	● ○ ○	● ● ●	● ● ●	● ● ○	● ○ ○
7) Haplotype compression performance	● ● ○	● ● ●	● ● ●	● ○ ○	● ○ ○
8) Ease of visualization	● ○ ○	● ● ○	● ● ○	● ● ●	● ● ●
9) Loci visualization and interpretability	● ○ ○	● ● ○	● ● ●	● ● ○	● ○ ○
10) Metadata and annotation	● ● ○	● ● ●	● ● ○	● ○ ○	● ○ ○
11) Compatibility with a linear reference coordinates	● ○ ○	● ● ●	● ● ●	● ● ○	● ○ ○

III. Construction : challenges

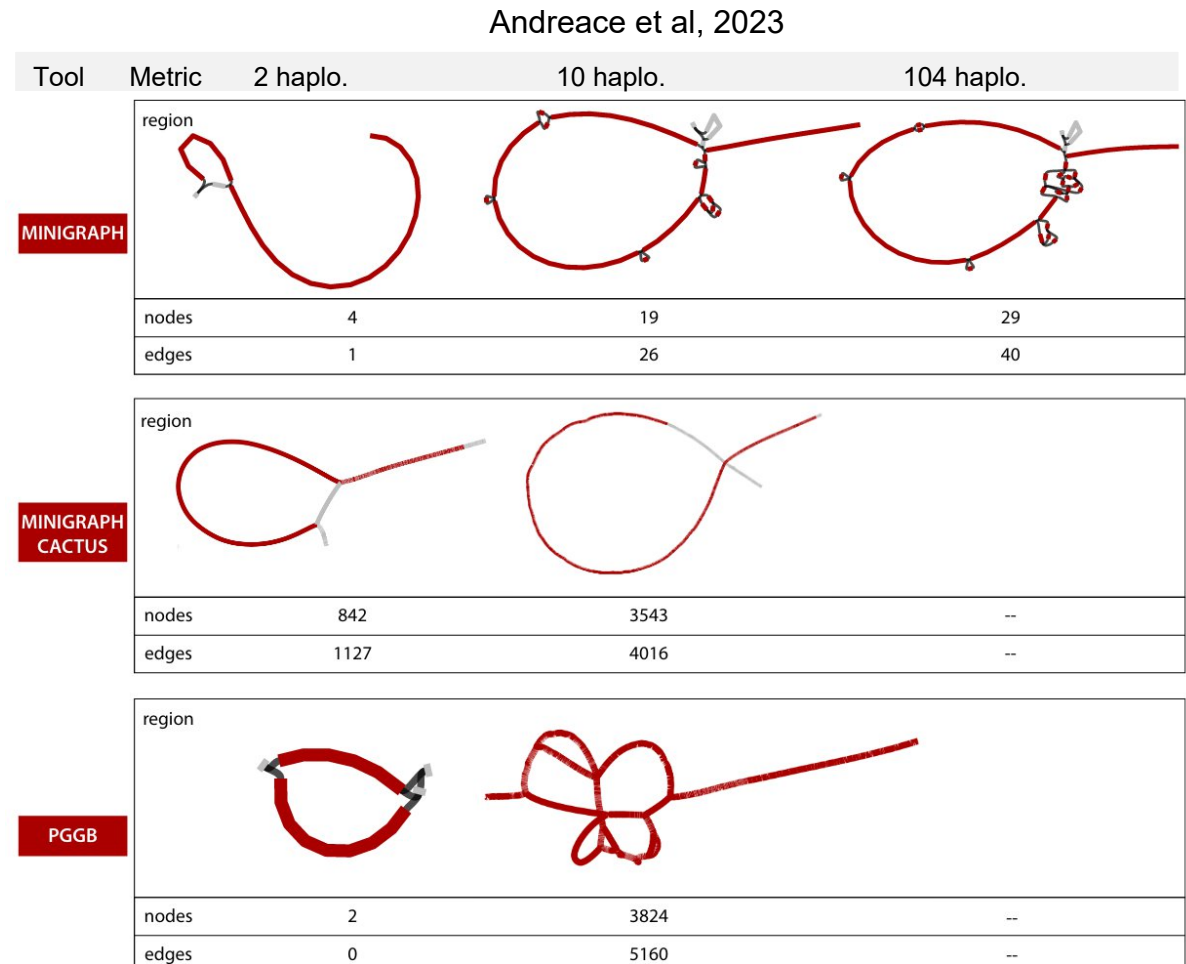
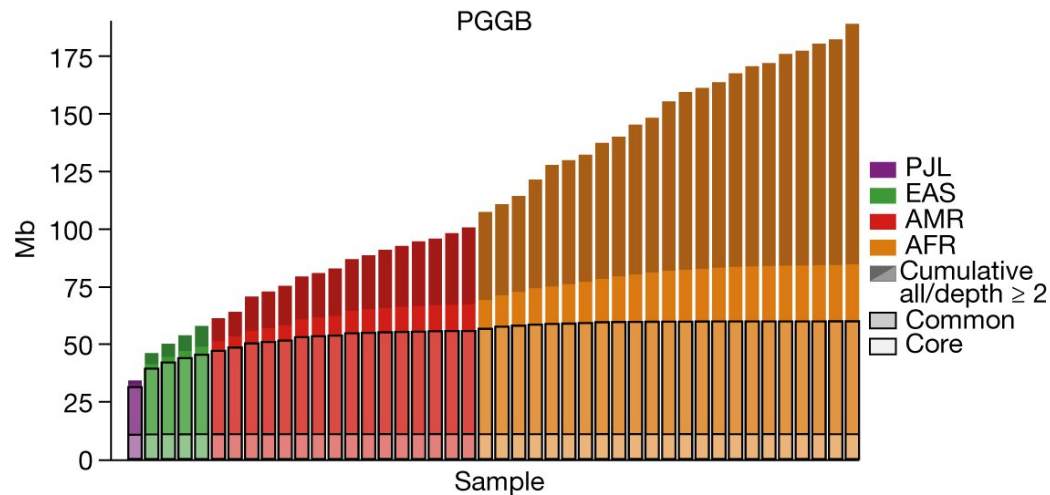
- Integrating every genomic variation from 10s to 100s of long genomes is not trivial !
 - Today 4 methods (from 2 research groups)
 - For 12 bovine assemblies (Leonard AS et al, 2023)
 - Graphs are huge !
 - Computational requirements can be huge !

Parameter	Unit	minigraph	cactus	pggp
Nodes	N	427,012	198,431,246	179,575,371
Edges	N	606,926	272,102,708	245,150,846
Node length	bp	2,598,811,581	3,041,026,095	3,012,039,323
Path steps	N	3,358,976	1,621,936,527	1,442,793,659
Repetitive sequence	bp	1,107,501,421	1,361,489,638	1,415,552,890
Centromeric sequence	bp	2,939,789	291,982,193	255,091,362
CPU time	h	14 ^a	226 ^b	3,559
Max memory	GiB	7	54	46
GFA file size	GB	2.6	26.1	23.7



III. Construction : challenges

- A balance : variation length / variant frequency / # integrable genomes
 - All methods discard “some” variations to reduce graph size or complexity.
 - 50 human haplotypes, graph size still continue to increase. (many rare variants)

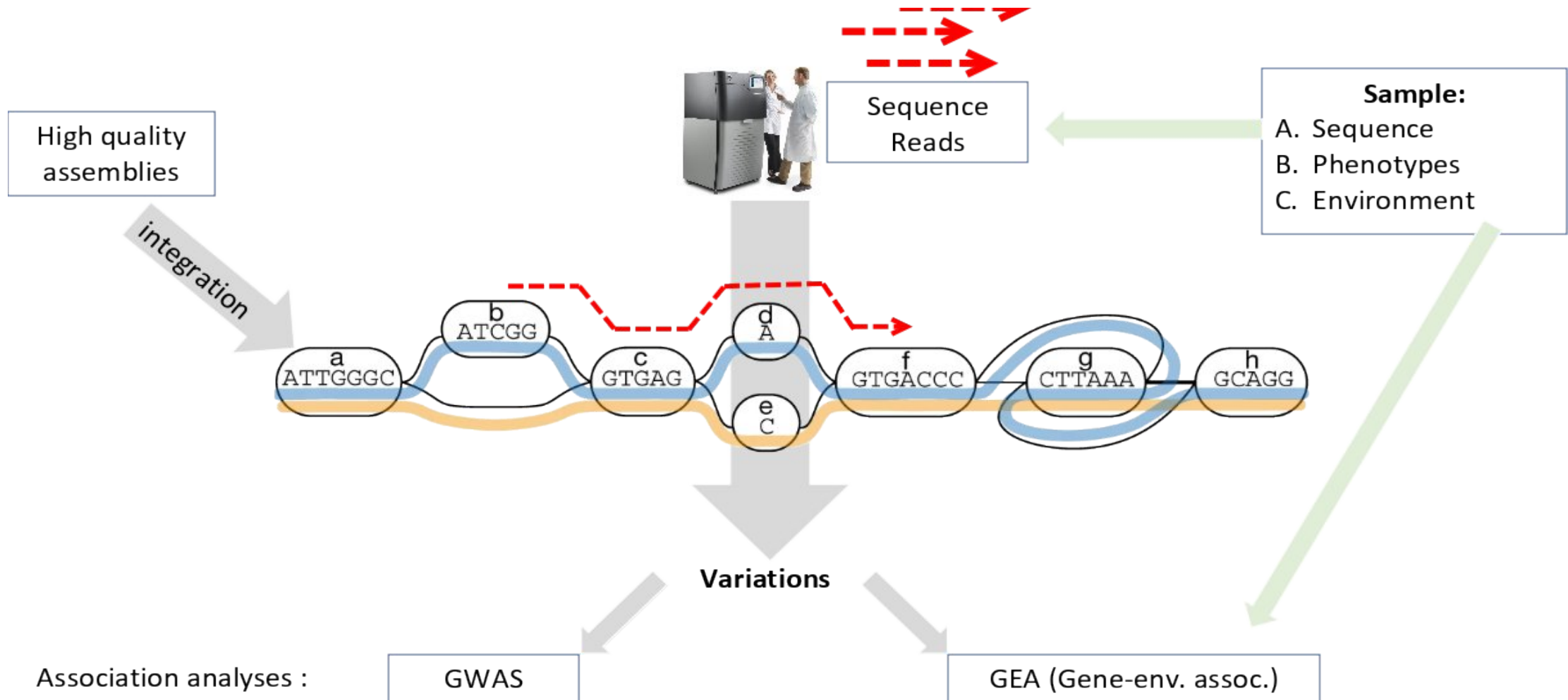


III. Construction

- ▶ Building a variation graph will not be smooth. Keep clam, this is expected.
- ▶ Methods and tools evolve rapidly and are not stable.
- ▶ Today: choose between Minigraph / Minigraph-Cactus depending on targeted scale.
- ▶ PGGB should improve soon...
- ▶ Many tools for graph postprocesses: impact poorly evaluated today.
- ▶ Pipelines are self sufficient, but hard to tune.
- ▶ Current default parameters were optimized for human. Impacts on non-primate metazoans, plants, fungi pangenome graphs ?

IV. Mapping and variant calling

- Today, main application is variant calling, genotyping.



IV. Mapping and variant calling

- “seed and extend” approach

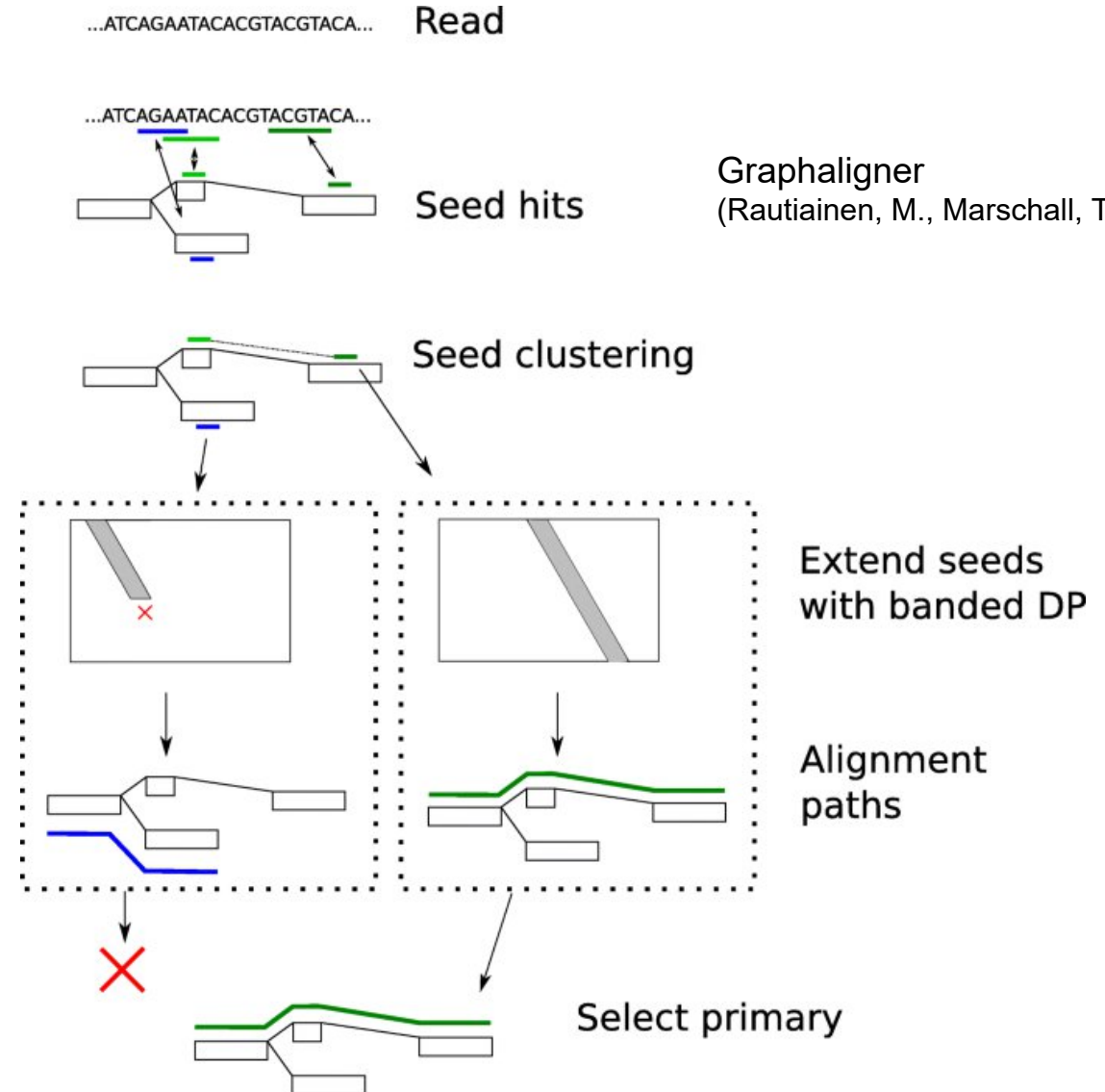
1. Pre-requisite => graph k-mers indexation (seeds indexation)

2. Mapping itself :

1. **Find anchors** between query and graph nodes (k-mer seeds)

2. **From these anchor, select a path in the graph**

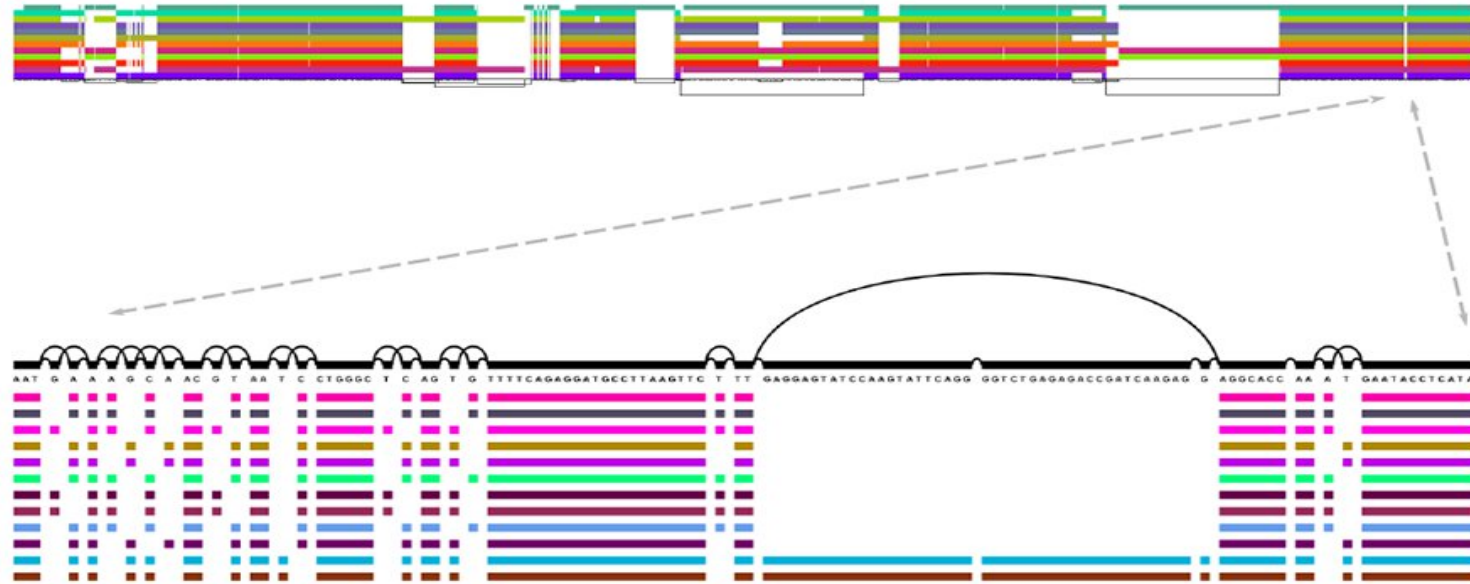
3. Then, do a “classic” **alignment** between the query sequence and selected path



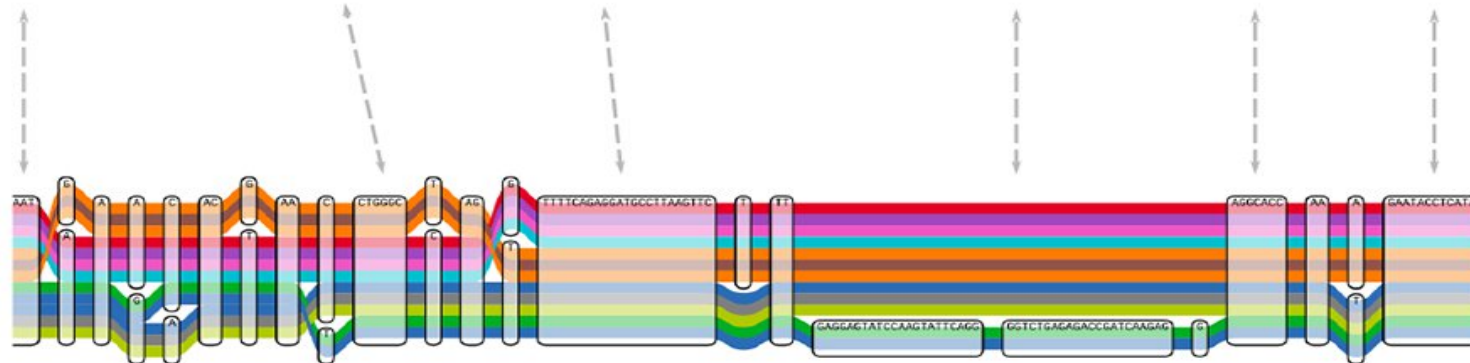
V. Visualisation

- Likely the most limited aspect of VG tools today :
 - How to switch from coordinate-based to graph-based representation ?
 - How to make sense of huge & complex topologies in 2D ?

ODGI
(Guarracino et al, · 2022)

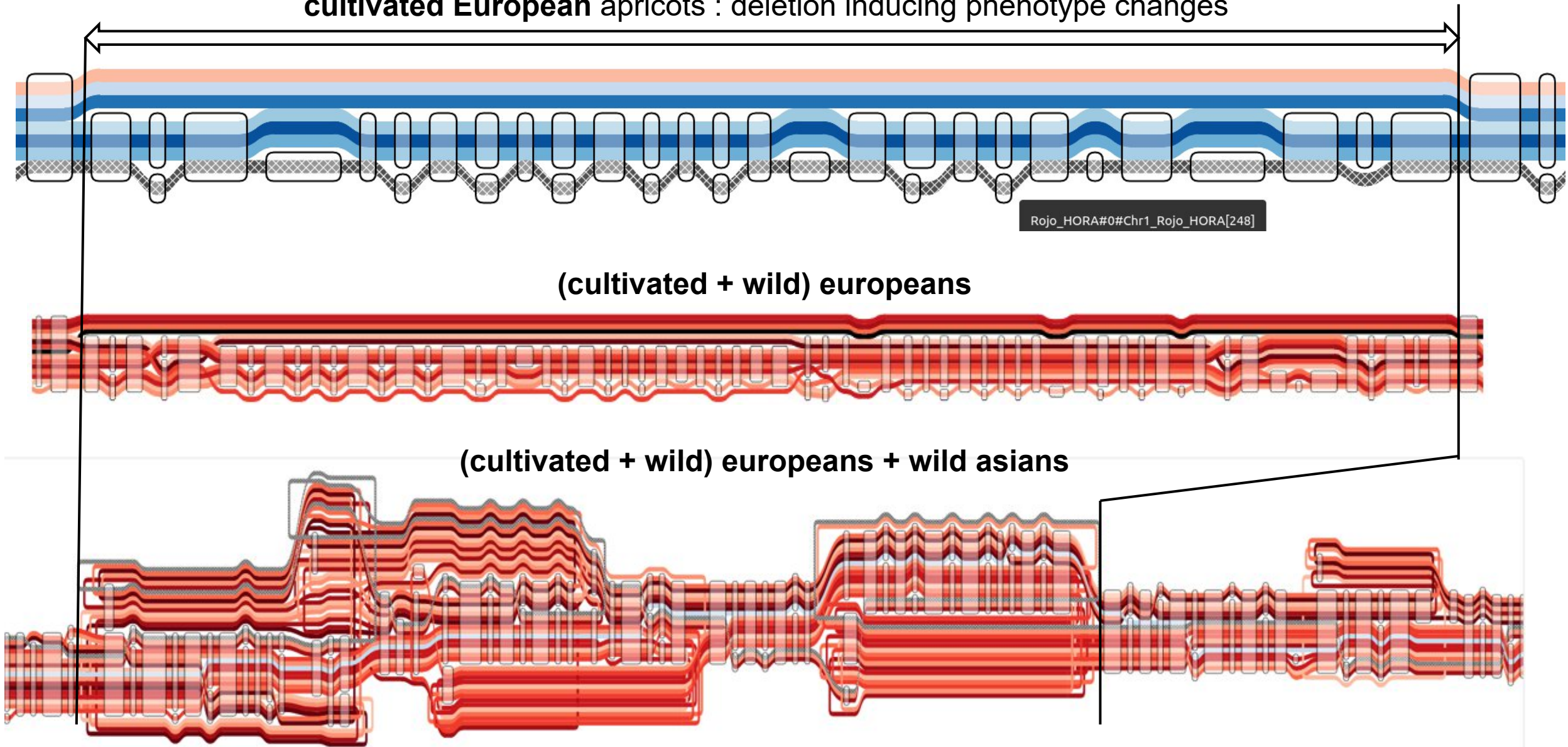


SequenceTubeMap
(unpublished)



V. Visualisation

cultivated European apricots : deletion inducing phenotype changes



Thank you for your attention.



A monthly-evolving pangenome graph. ;) (<https://www.pangenome.eu/>)