Introduction to pangenome graphs

Benjamin Linard

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

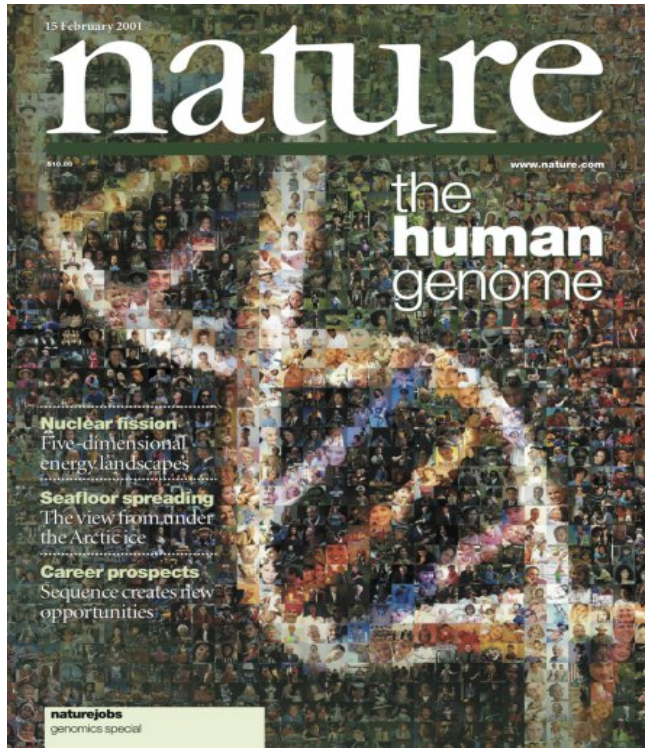
06-06-2024

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

MIA TOULOUSE

I. Contexte



2001

Premiers génomes « complets »

2010

Premiers aperçus de la diversité des génomes

May 2023

Changement de paradigme. Contextualiser l'analyse bioinformatique avec la totalité de la diversité répertoriée

I. Contexte : génomes (enfin) complets

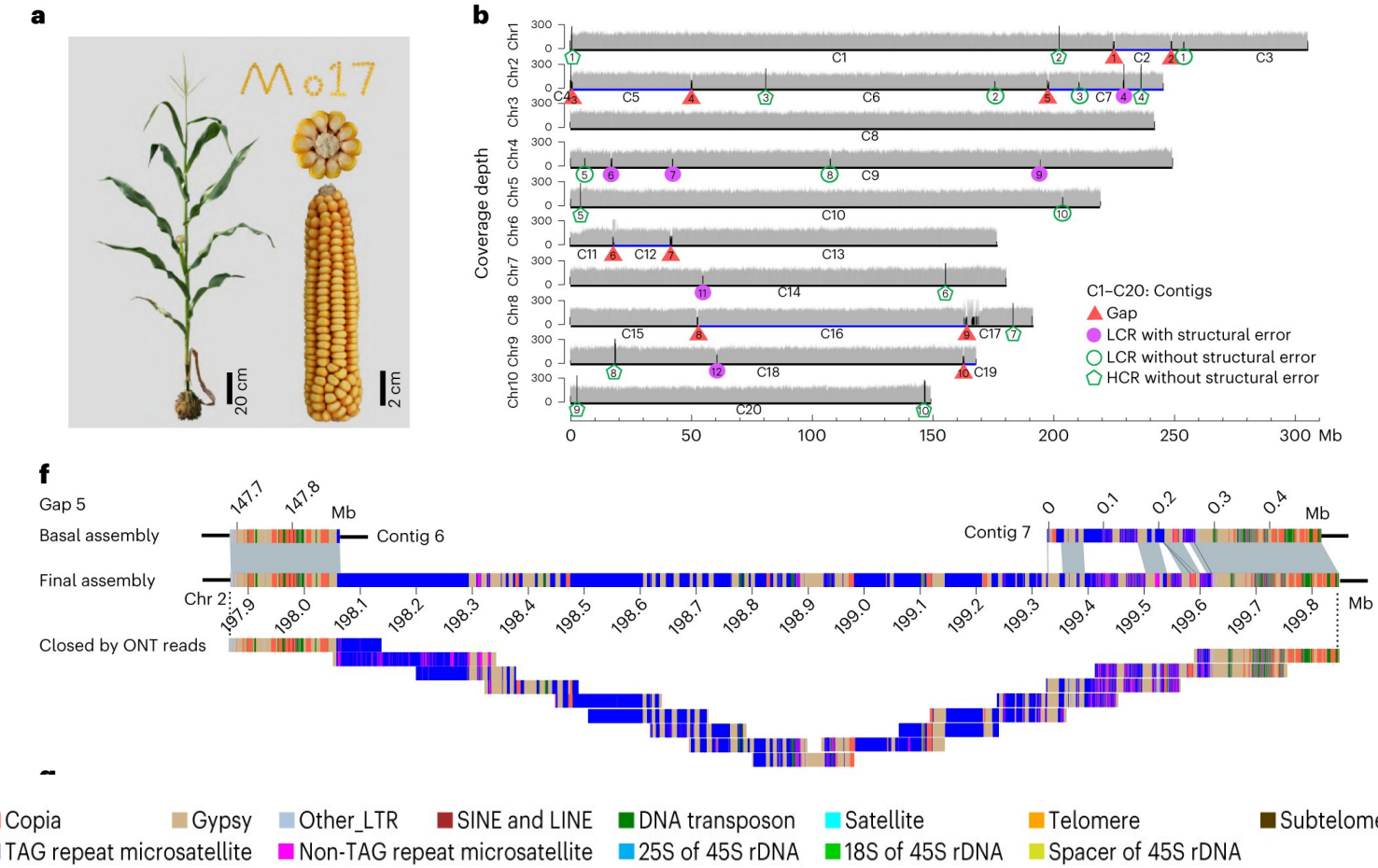
Why now ?

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 !

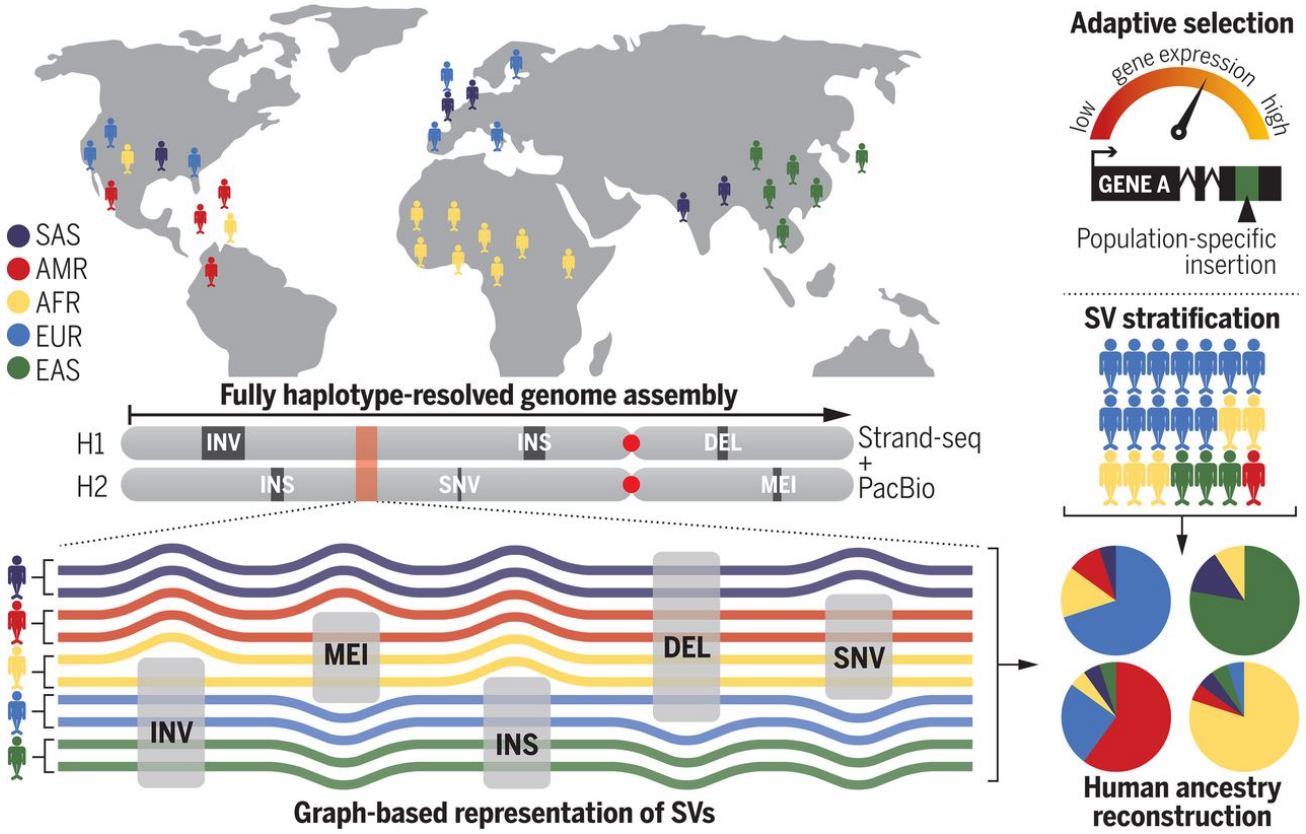
Chen et al, june 2023.
"The 2,178.6 Mb T2T Mo17 genome with a base accuracy of over 99.99% unveiled the structural features of all repetitive regions of the genome."



I. Contexte : de la référence unique à la population

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

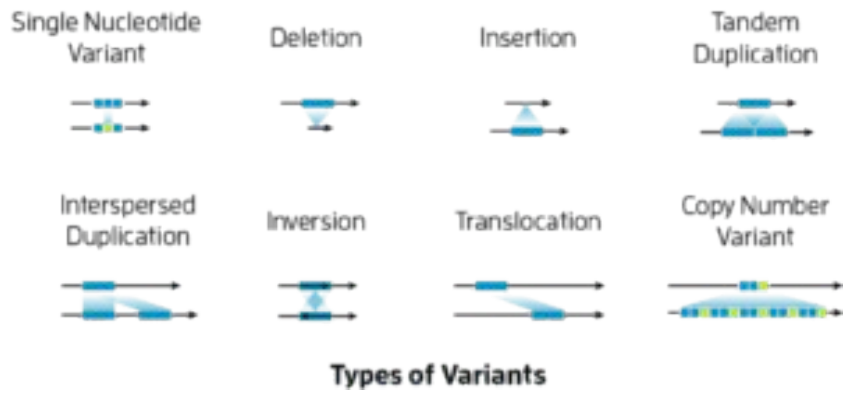
- *"The current version of the reference genome (GRCh38) is estimated to miss up to 10% of our species genetic information" (SS Sherman RM, 2020)*



P Ebert et al. *Science*, 2021

I. Contexte : variants structuraux (SVs)

- **A major novelty :**
  full access to genome
  Structural variants (SV)

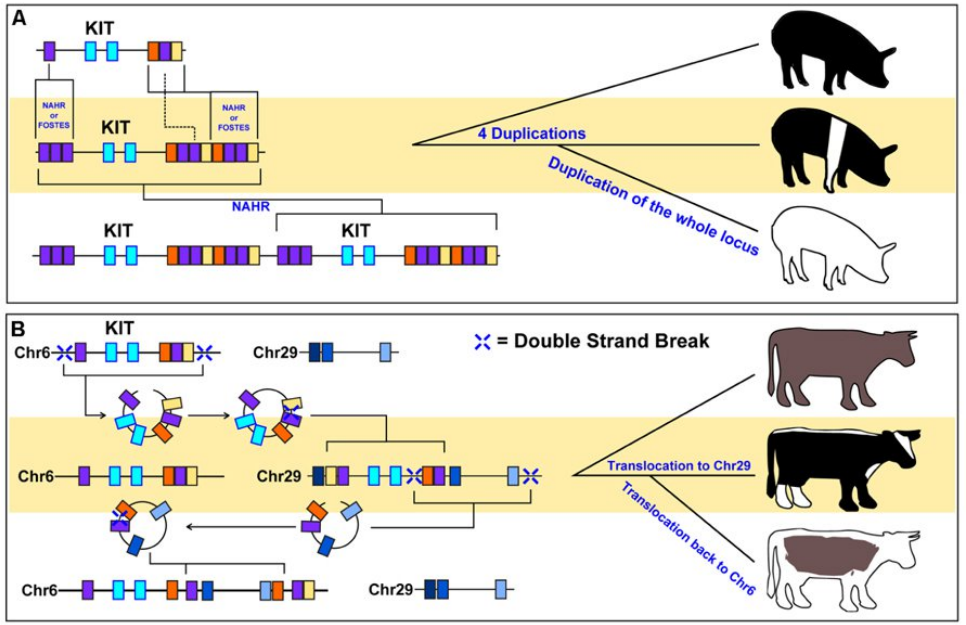


- Unexplored impact of SVs

Dosage effects

Positionnal effects

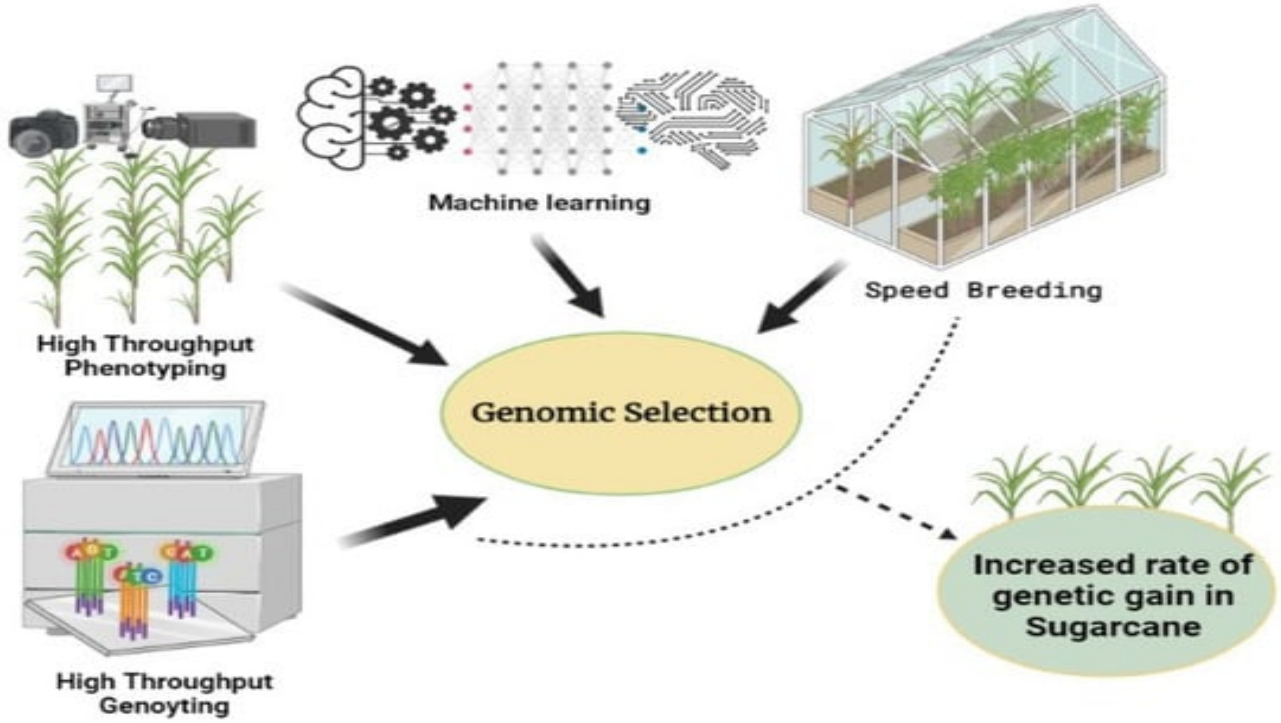
→ **Disruptions: Gene or regulation** →



Bickhart DM & Liu GE, 2014

I. Contexte : agronomie et sélection

- **Nouveau contexte de génération des données :**
  - Technologies HiFi et Hi-C
  - Génomes complets « télomère à télomère »
  - Assemblage routinier
- **Nouvelles opportunités pour la génétique :**



Sandhu et al, 2022. Plants

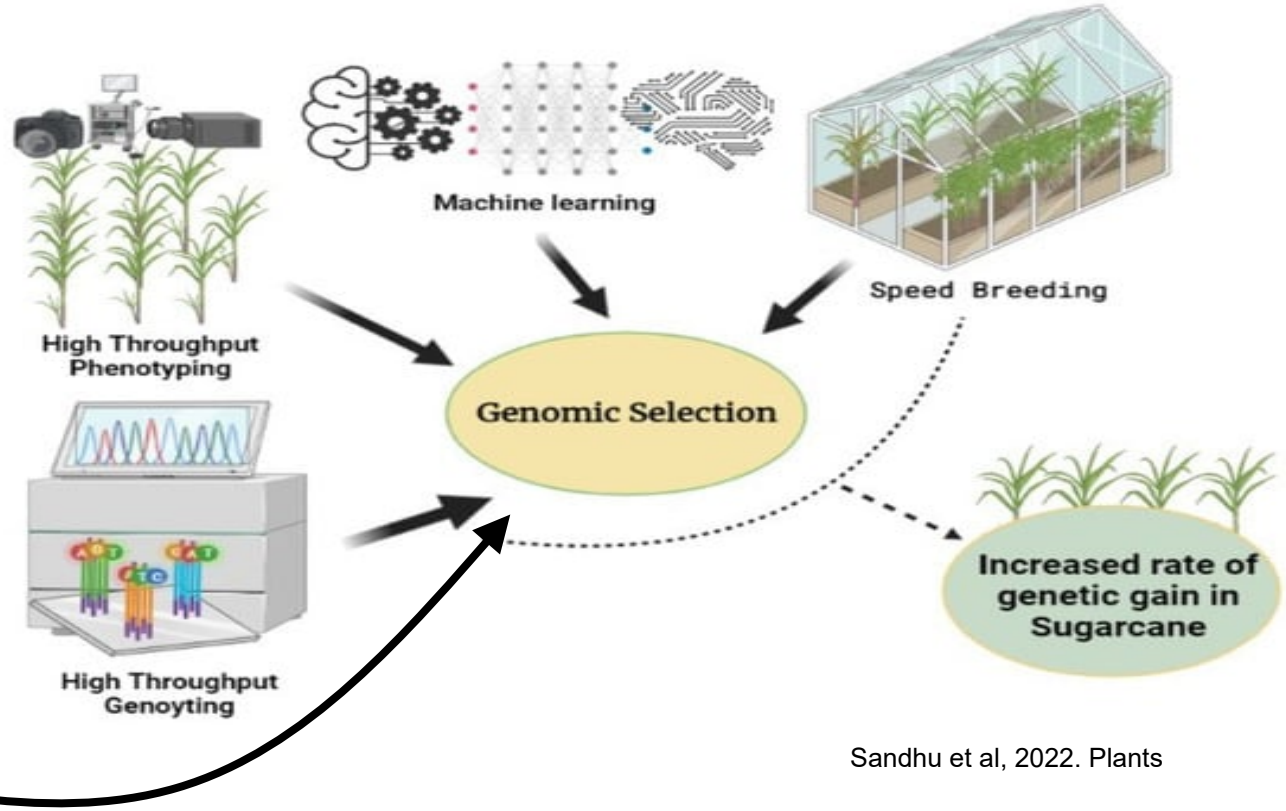
I. Contexte : agronomie et sélection

- **Nouveau contexte de génération des données :**
  - Technologies HiFi et Hi-C
  - Génomes complets « télomère à télomère »
  - Assemblage routinier

- **Nouvelles opportunités pour la génétique :**

  - Génomes **complets**
    => variants structuraux
  - **Nombreux** individus
    => fréquence des variants

Diversité génomique de l'espèce



Sandhu et al, 2022. Plants

II. Modèles

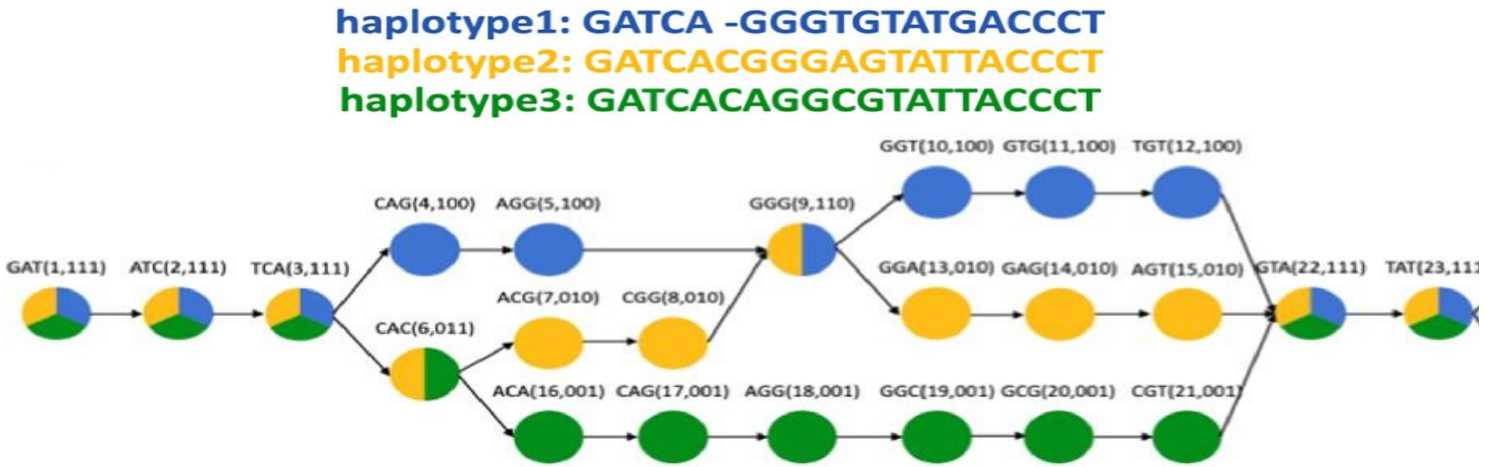
- **Modéliser la diversité des génomes par un graphe**

Graphes de
De Bruijn

compactés / colorés

Microbiomes
(majoritairement)

(>100s petits génomes)

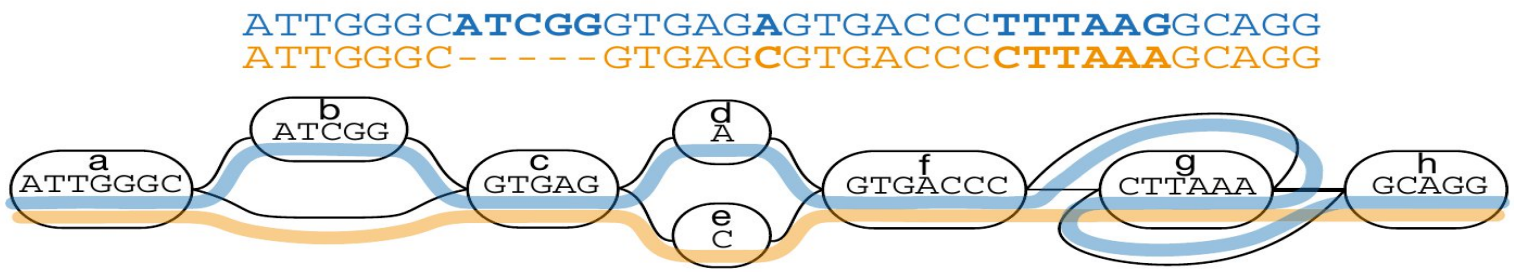


Guo, J. *et al.* (2021)

Graphes de variation

Génomes eucaryotes

(10s à 100s de longs génomes)



II. Modèle VG : limiter les biais d'analyse

(a)

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

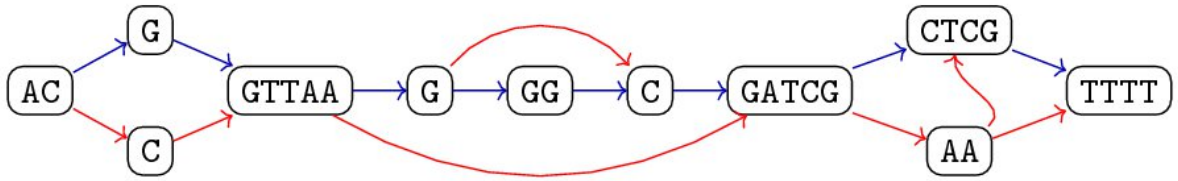
Reads:

```
ACCGTTAAGCGA
TCGAATTTT
```

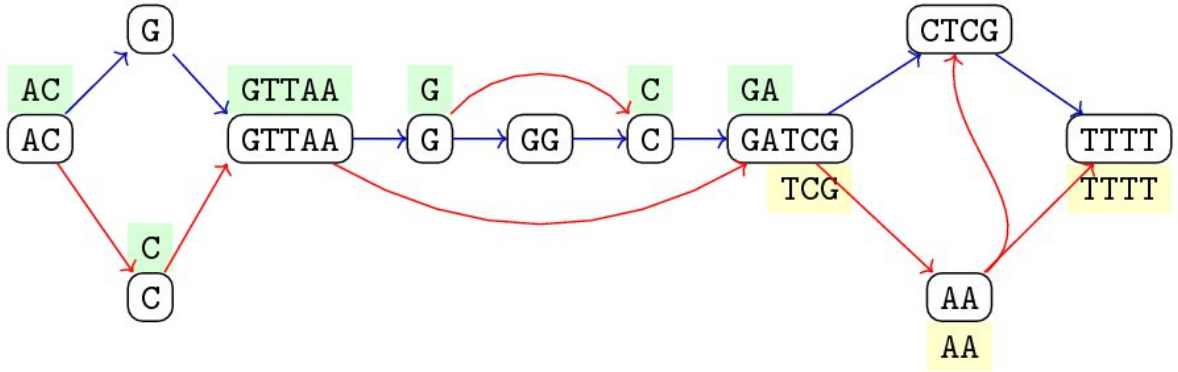
(c)

```
ACCGTTAAGCGA
ACGGTTAAGGGCGATCGCTCGTTTT
            TCGAA--TTTT
```

(b)

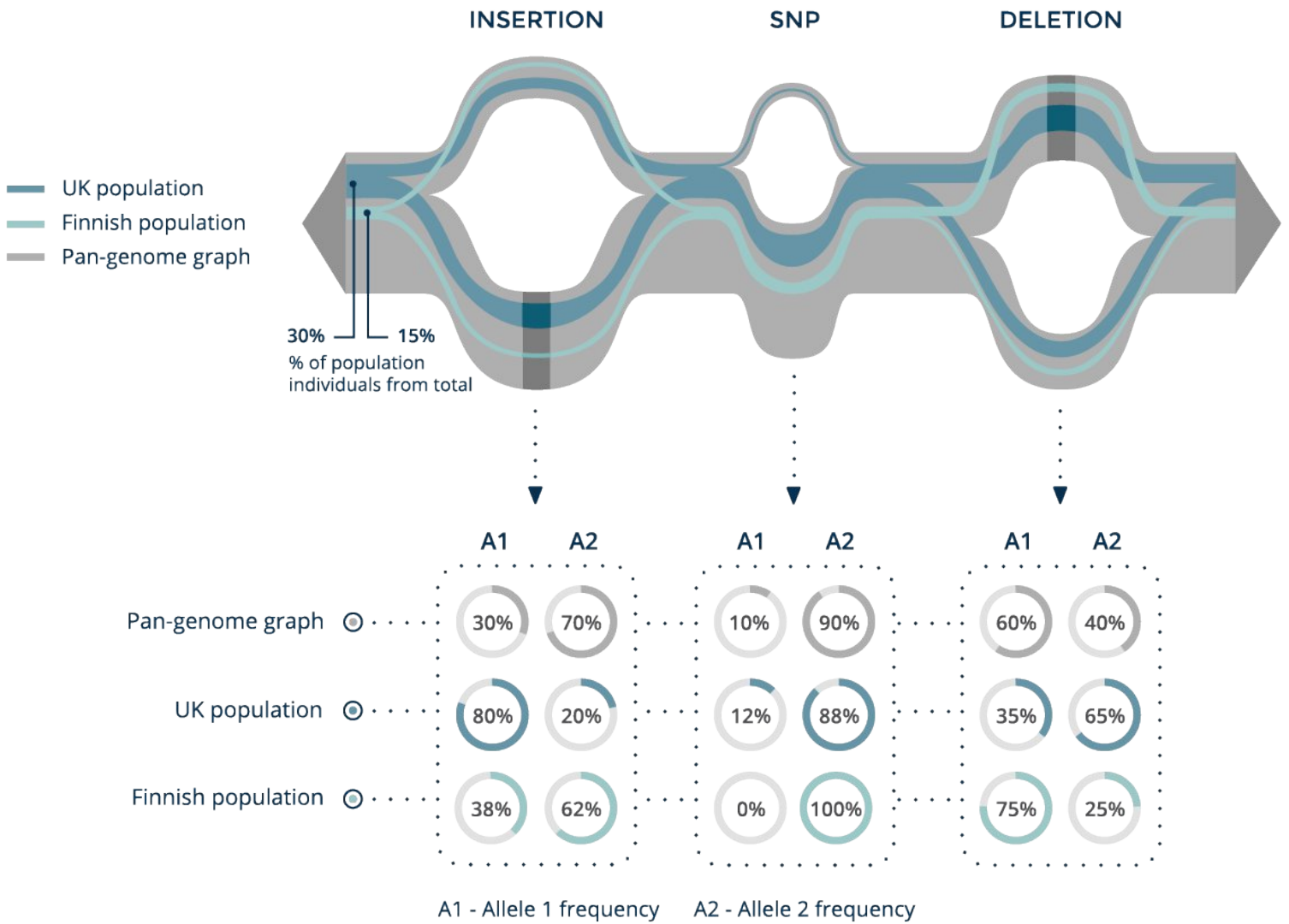


(d)



JA Baaijens et al, 2022

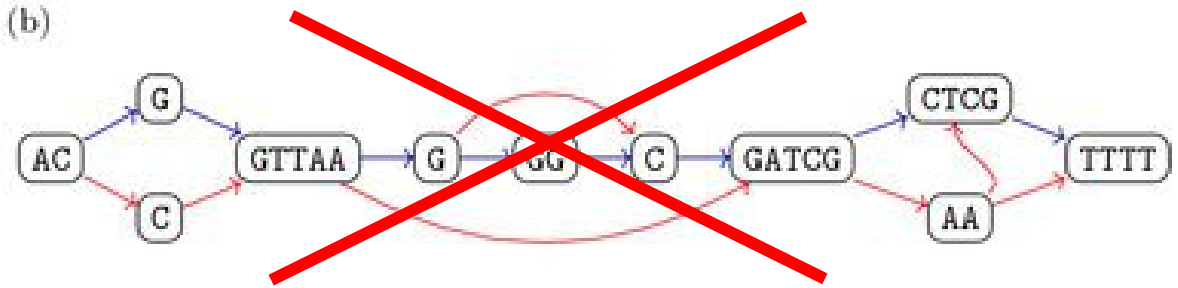
II. Modèle VG : limiter les biais d'analyse



www.sevenbridges.com/graf

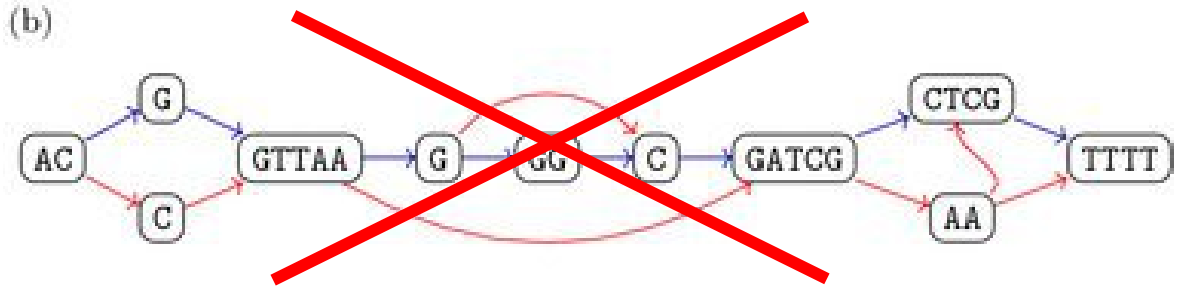
II. Modèle VG

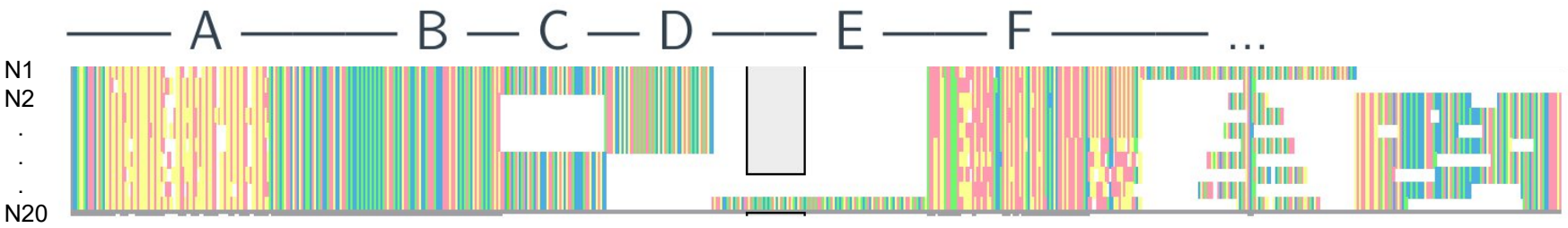
- **En pratique ...**

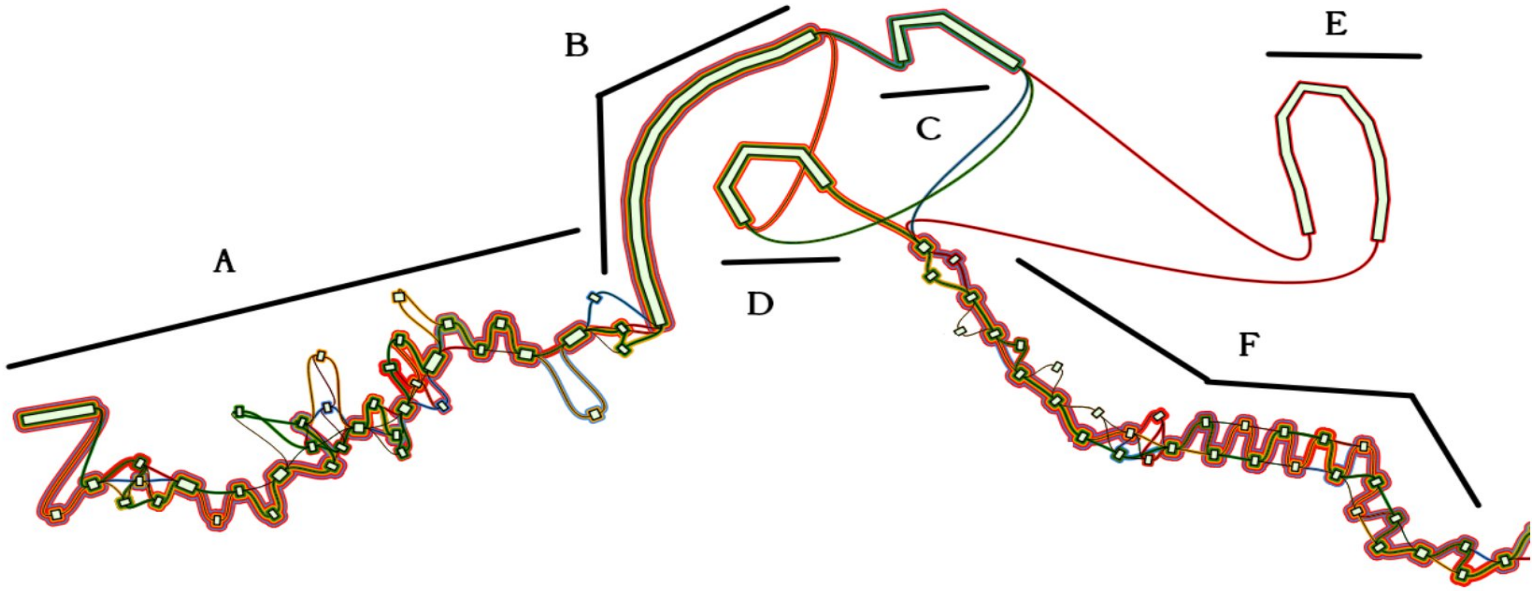


II. Modèle VG

- **En pratique ...**

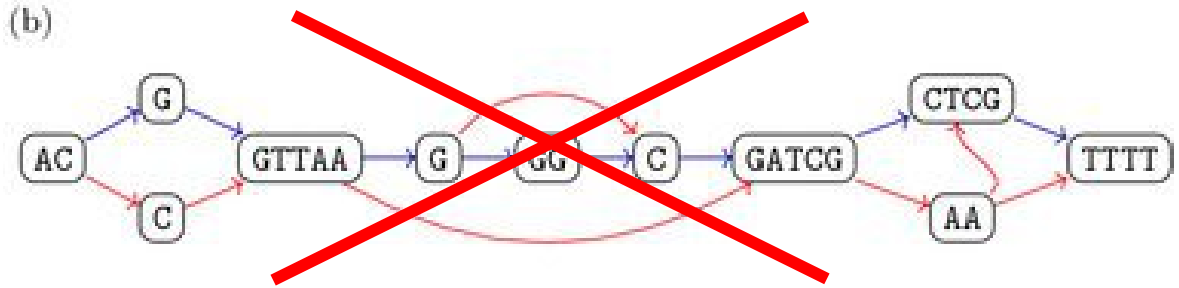


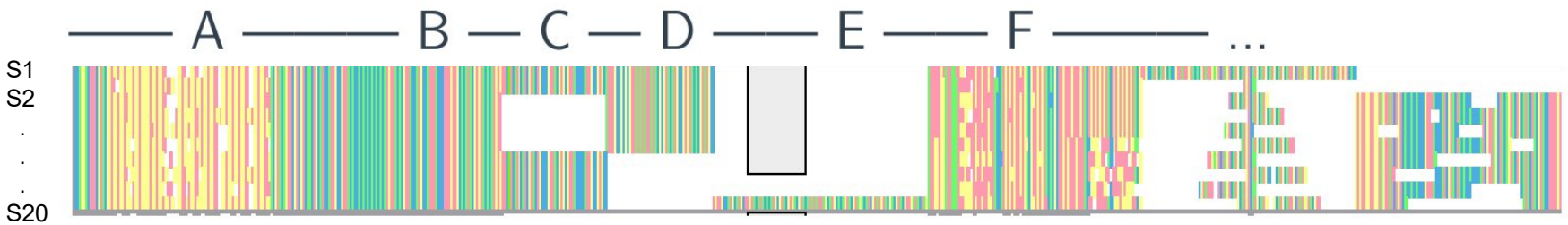


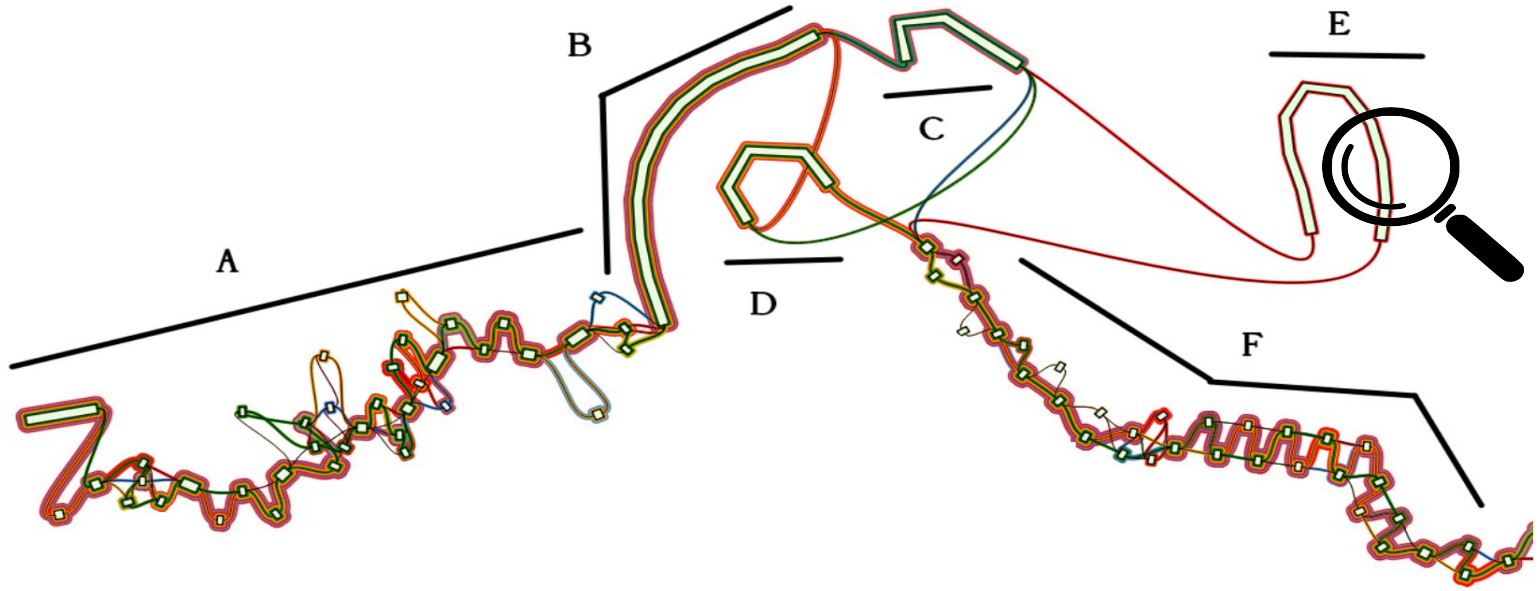


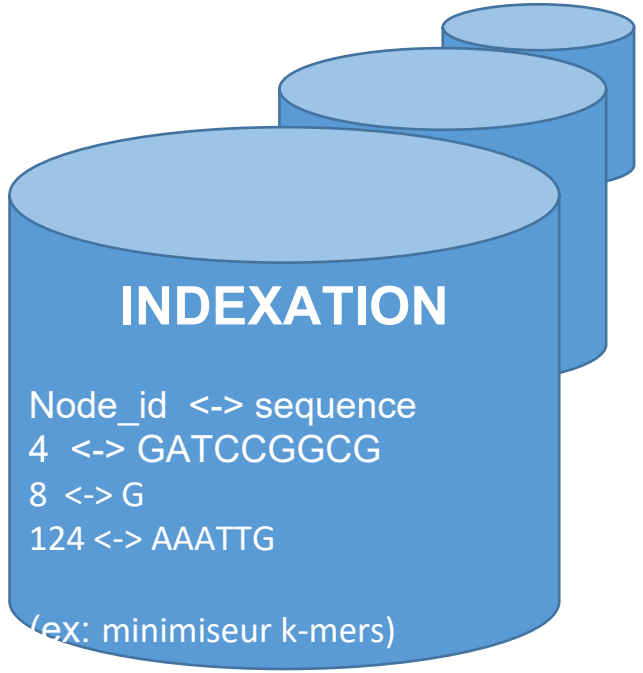
II. Modèle VG

- **En pratique ...**



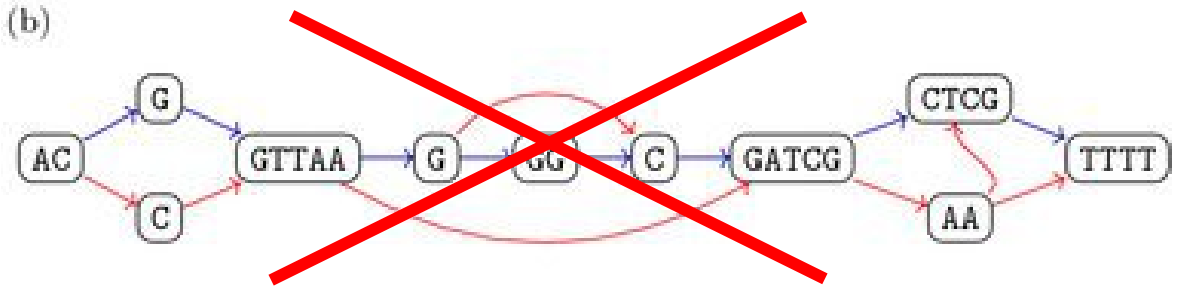


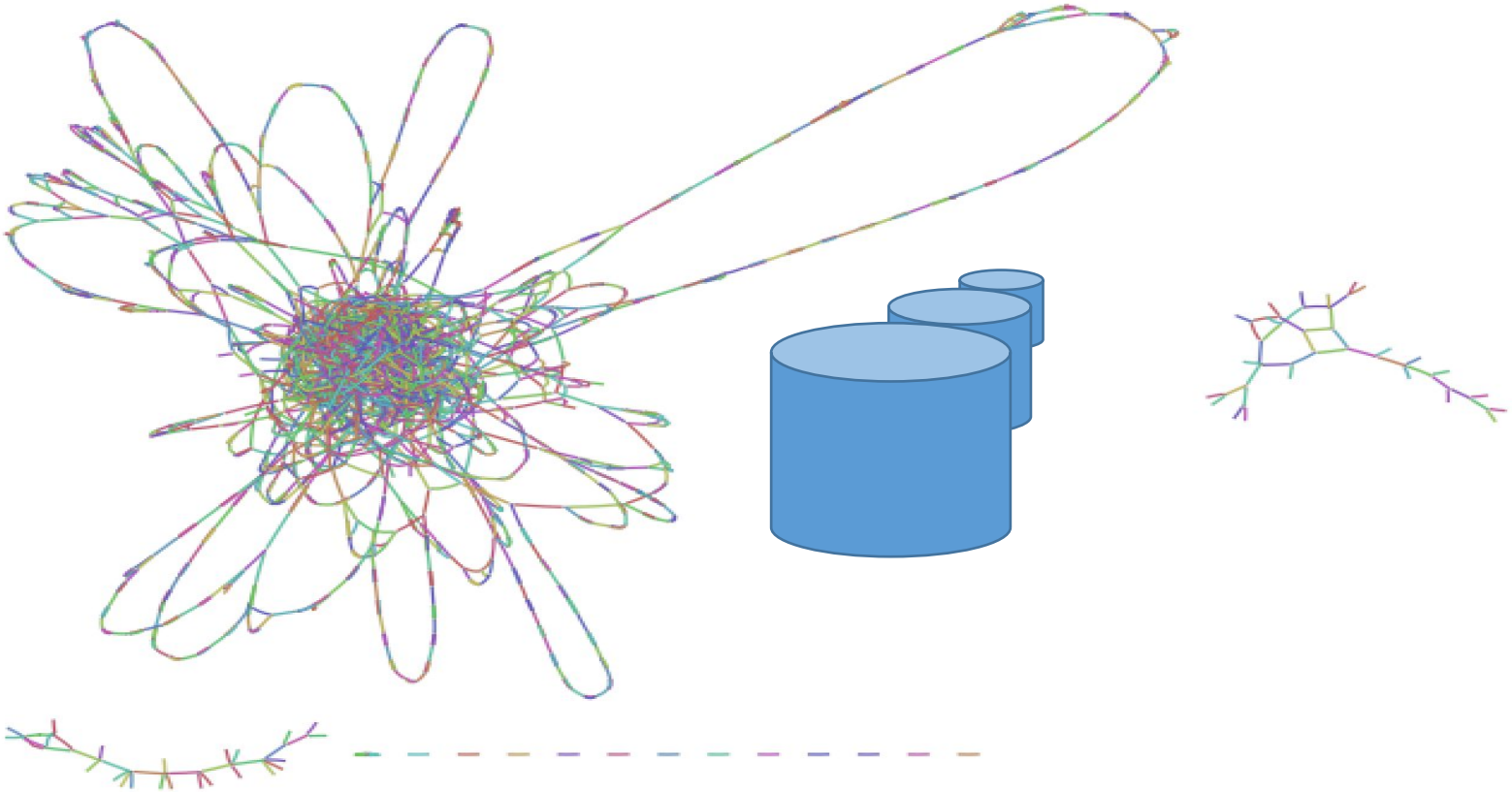




II. Modèle VG

- **En pratique ...**





II. Modèle VG : les défis

- **Défis liés au modèle et son application**

Dimensions du graphe :

- 12 assemblages bovins, $>10^7$ nœuds (Leonard AS et al, 2023)
- inadapté à la recherches par parcours
- défis liés à la mémoire et la taille des index
- nœud : associé de 1 à *n* kilobases

Topologie :

- nœuds de très haut degré
- nombre de chemins possibles
- cycles (beaucoup d'algos de recherche ne sont plus polynomiaux)
- séquence et fréquence associées aux nœuds: non classique en théorie des graphes (classique = pondération)

Applications :

- Interroger la diversité génétique : à quelle échelle ?
- Vers une bioinformatique basée sur les graphes et non les séquences ?

.

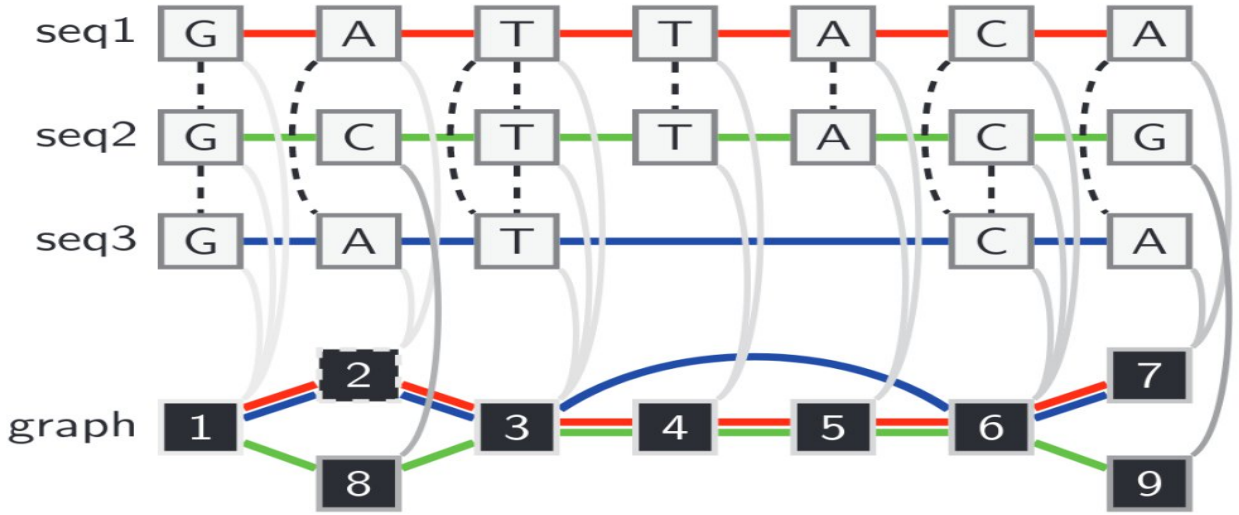
III. Construction : approche globale

- **Construction d'un graphe de variation**

  - Entrées: N un ensemble de génomes, alignés par paires
    A l'ensemble des homologies de paire de caractères

  - Sortie : Un graphe orienté $G = \{V, E, P\}$
    avec P , l'ensemble de chemins décrivant les génomes intégrés

  - Objectif : G inclut N et la totalité des relations par paires décrites par les alignements (G ne peut contenir moins d'information que dans A)



Garrisson et al, 2023

paths
seq1 : [1,2,3,4,5,6,7]
seq2 : [1,8,3,4,5,6,7]
seq3: [1,2,3,6,7]

III. Construction : approches disponibles

In its own category:

- Variation Graph (VG) (Garrisson et al, 2018)

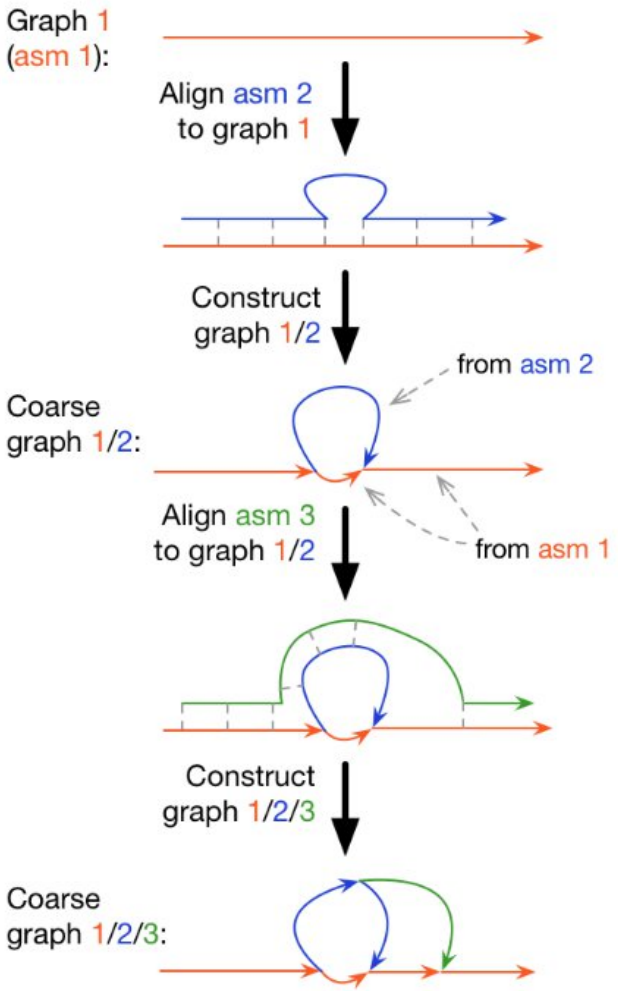
Genome alignment, graph construction, Post-processes.

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

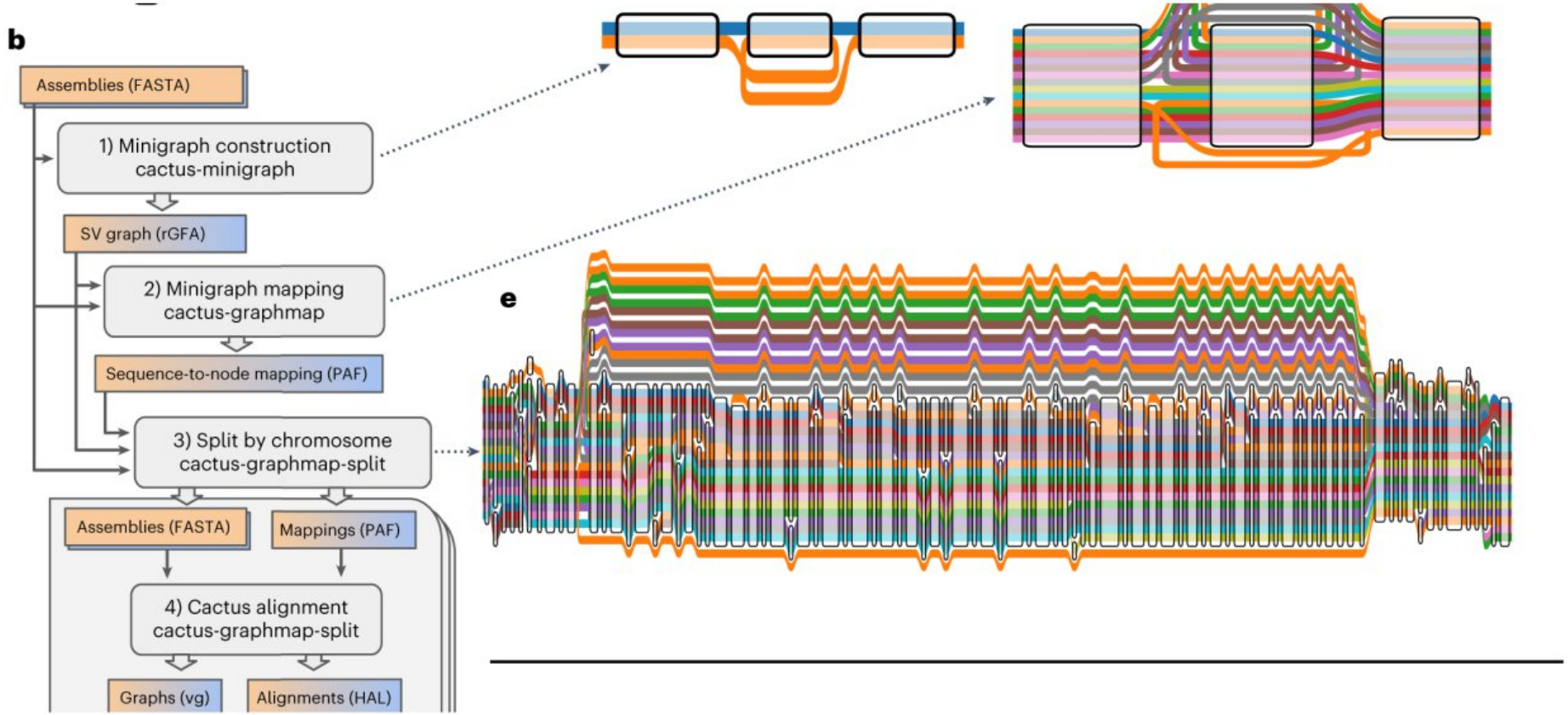


- INPUT:
  - A complete reference genome (fasta)
  - A set of other haplotypes (fasta)
- INCREMENTAL CONSTRUCTION:
  - Haplo 2 mapped to ref: output $\rightarrow G_{[ref,2]}$
  - Haplo 3 mapped to $G_{[ref,2]}$: output $\rightarrow G_{[ref,2,3]}$
  - ... etc ...

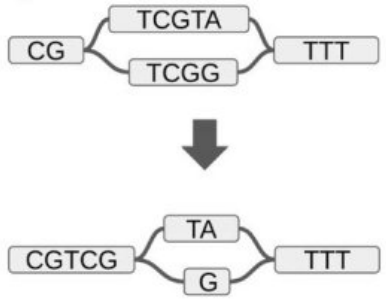
III. Construction : Minigraph-cactus

- INPUT:
  - At least 1, complete, high-quality reference genome (fasta)
  - A set of assemblies for other individuals (fasta)
  - (Optionnal) 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 $\rightarrow$ 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)

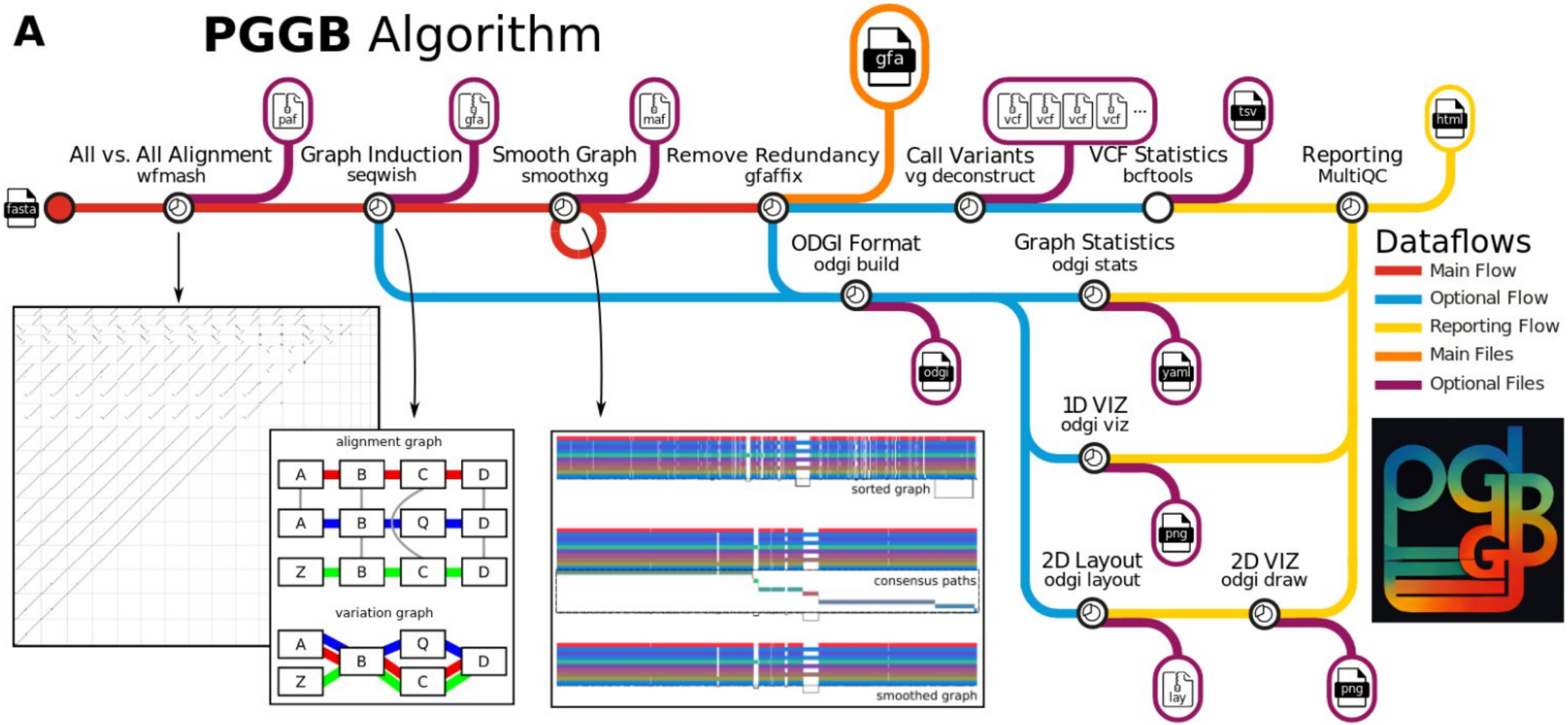
III. Construction : Minigraph-cactus



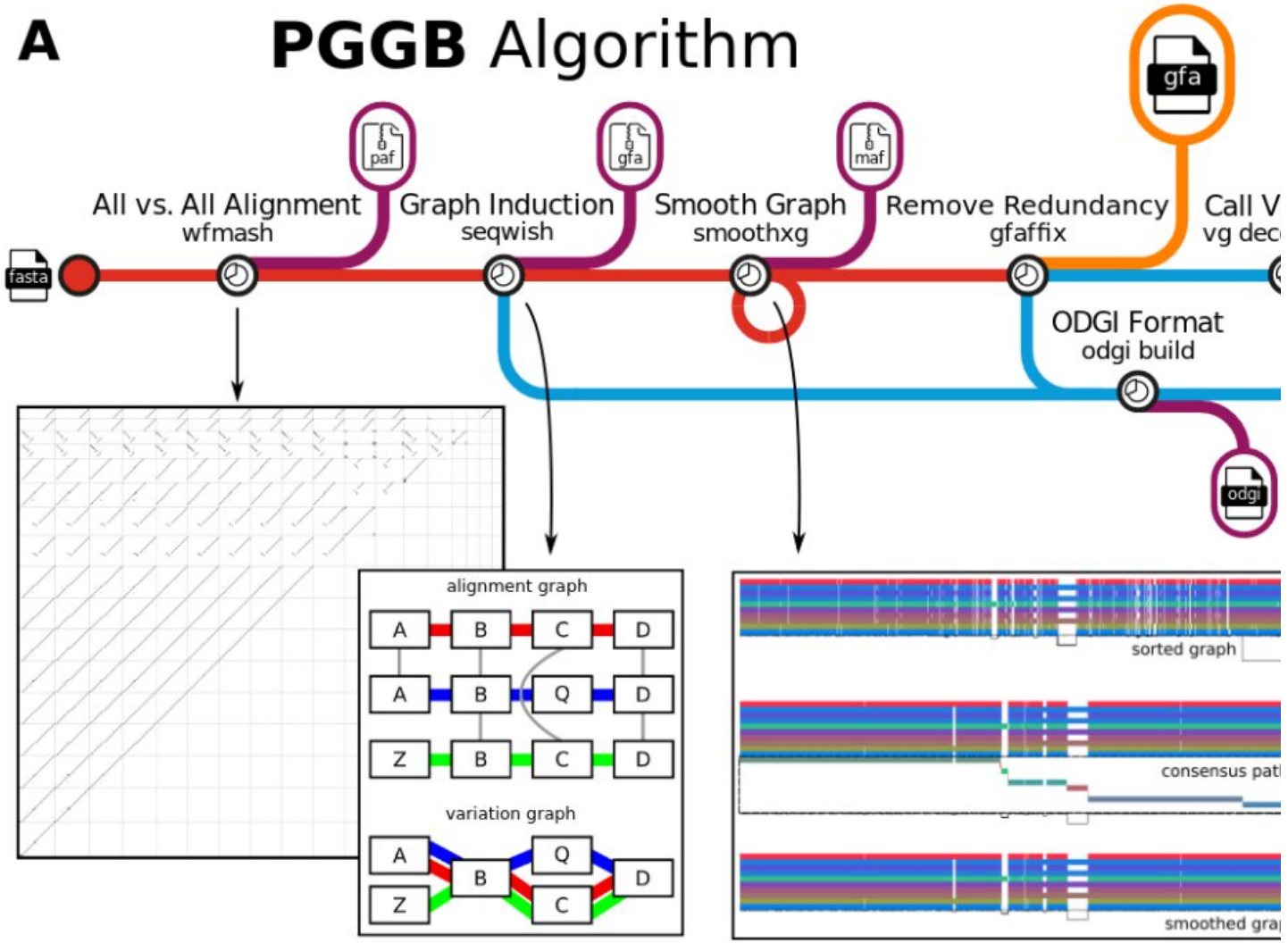
Post-process example: GFAffix tool



III. Construction : PGGB



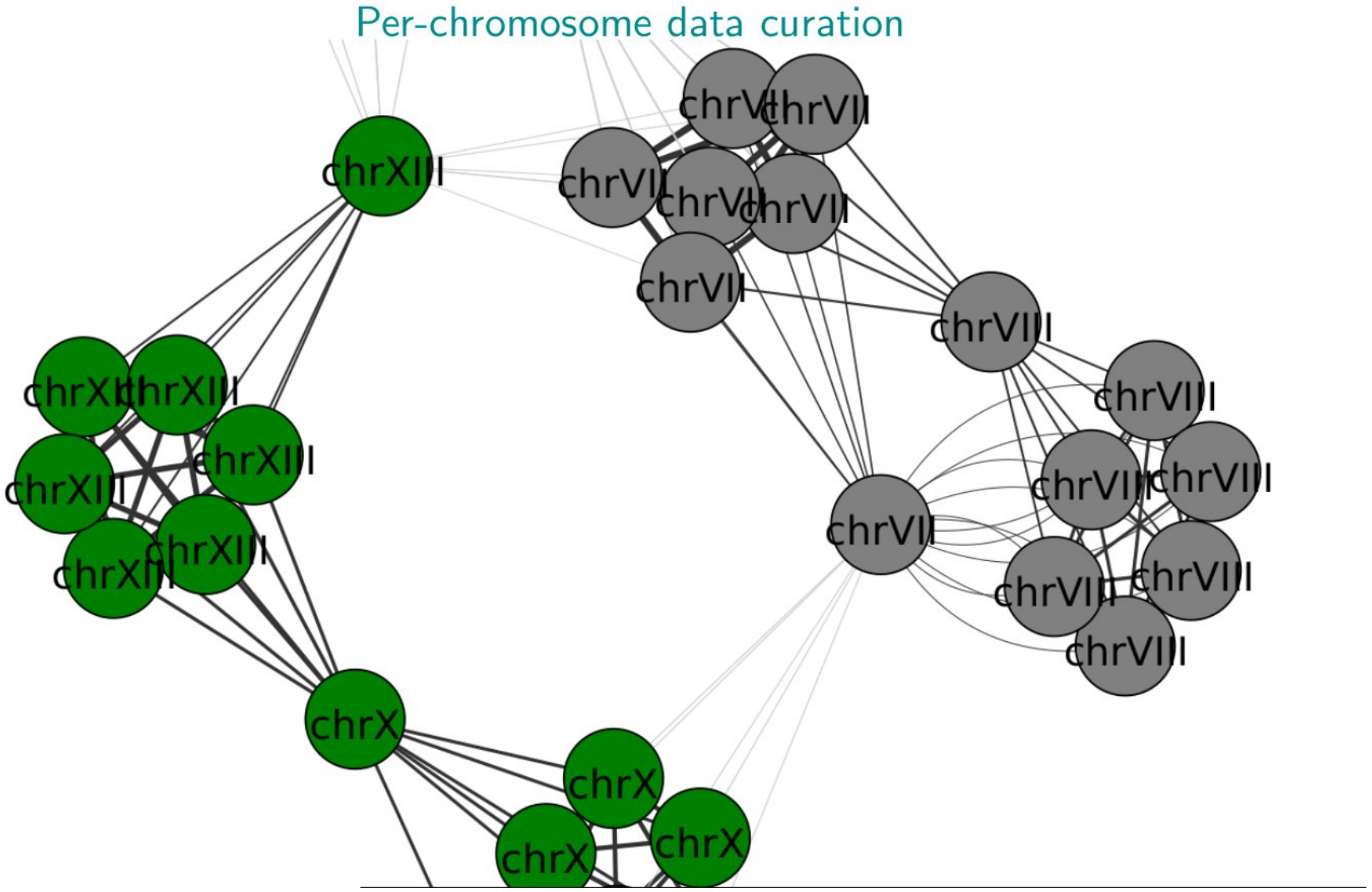
III. Construction : PGGB



III. Construction : PGGB

- INPUT:
  - A set of assemblies, one per chromosome.
- PROCESSUS:
  1. Paiwise 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.

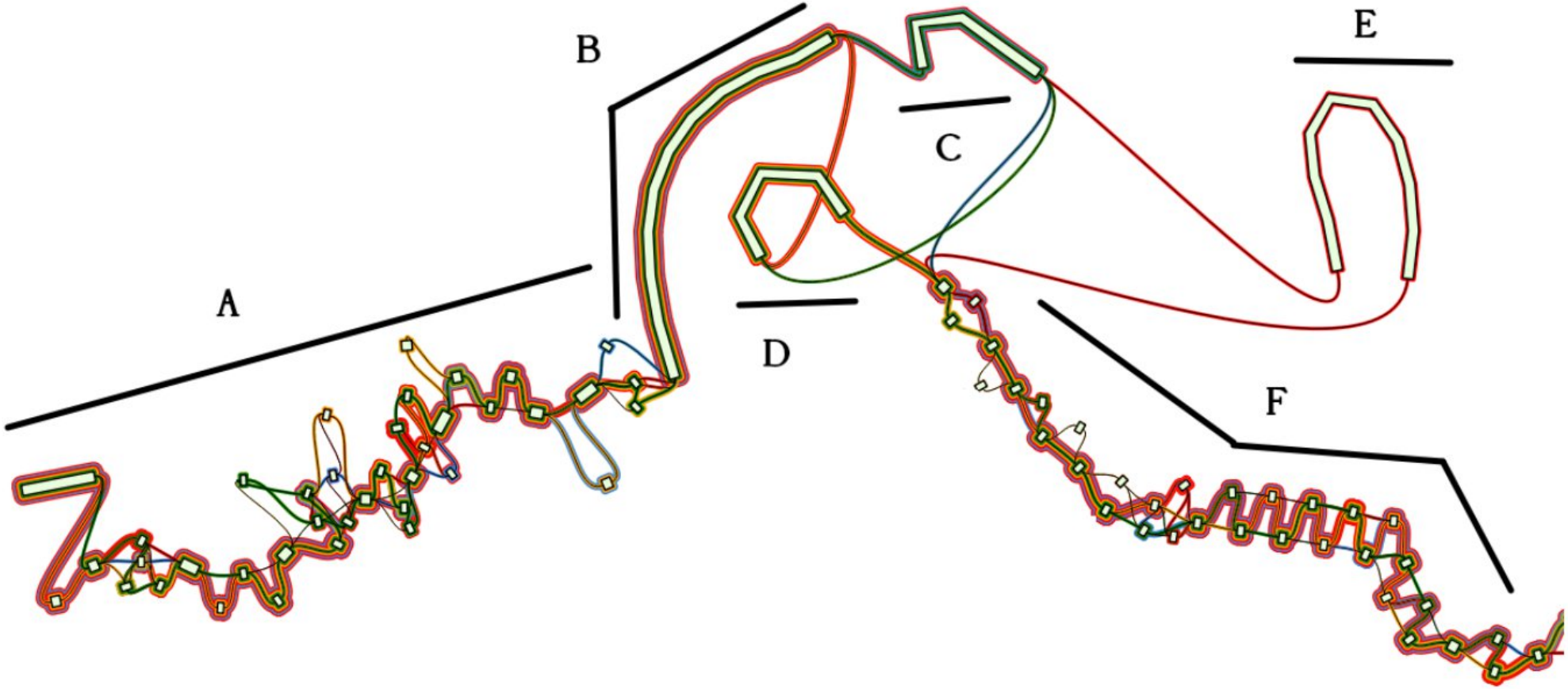
III. Construction : PGGB



III. Construction : PGGB

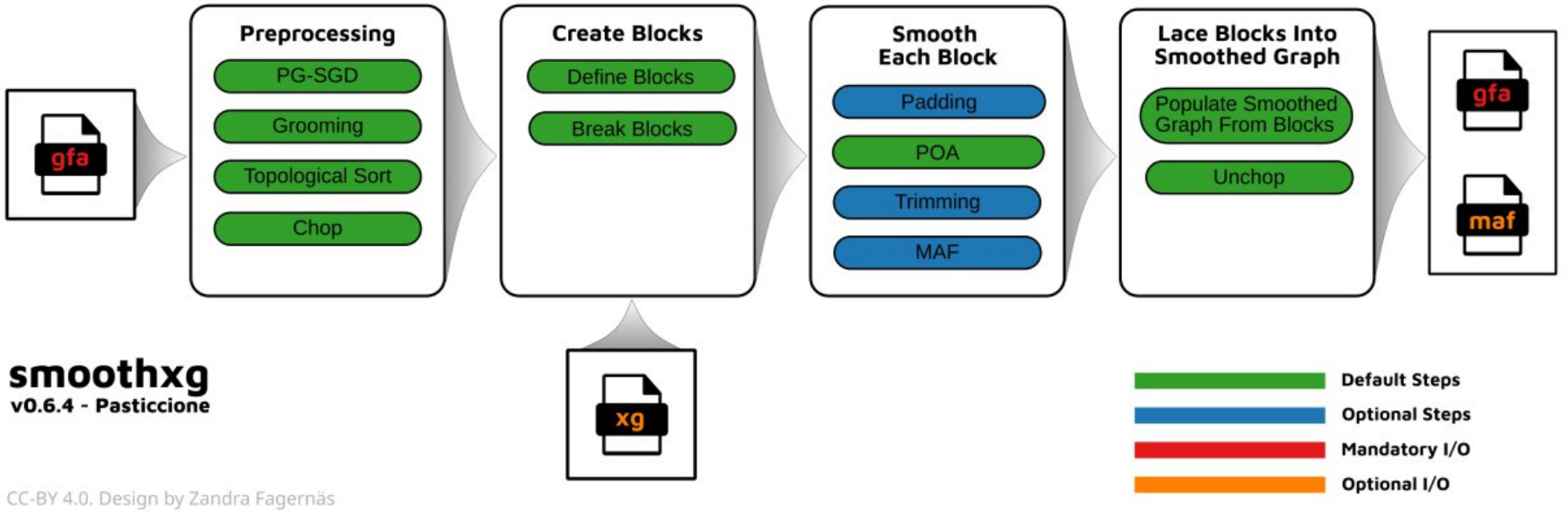
Seqwish output IS the expected graph

But this is NOT
the final PGGB graph.



III. Construction : PGGB

Recent paragraph + schematic in the doc, not much more in the PGGB paper...



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

III. Construction : différence fondamentale

Impact of genome order

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

III. Construction

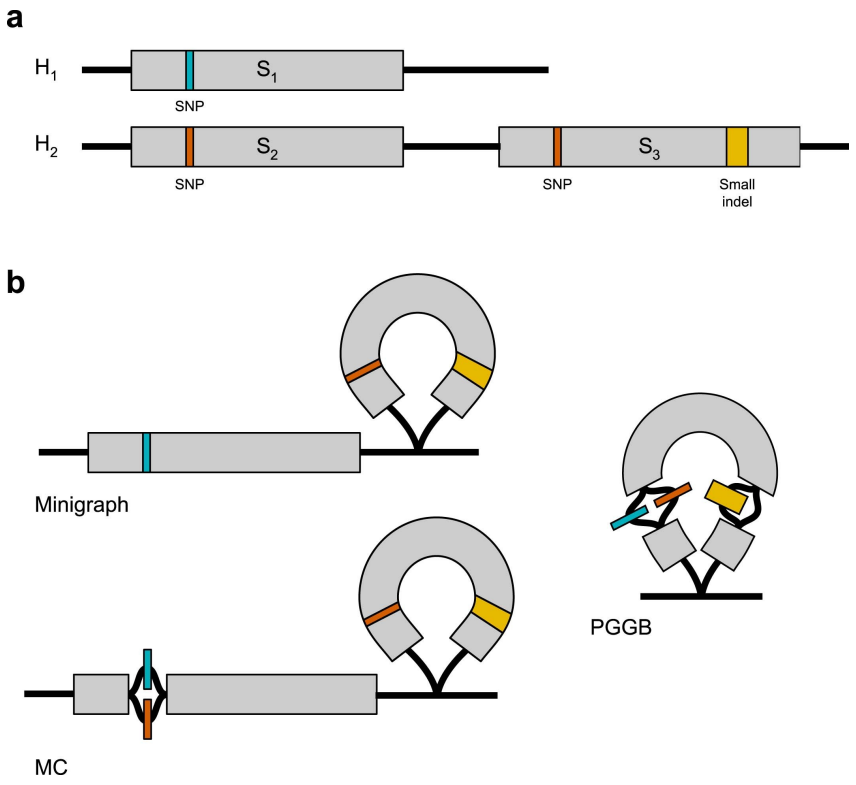
- Andreace et al 2023, https://doi.org/10.1186/s13059-023-03098-2

Metric	Bifrost	pggb	Minigraph-Cactus	Minigraph	mdbg
1) Construction speed	●●○	●○○	●○○	●●○	●●●
2) Variations	●●●	●●●	●●●	●●○	●●○
3) Scalablilty	●●●	●○○	●○○	●●○	●●●
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 : défis

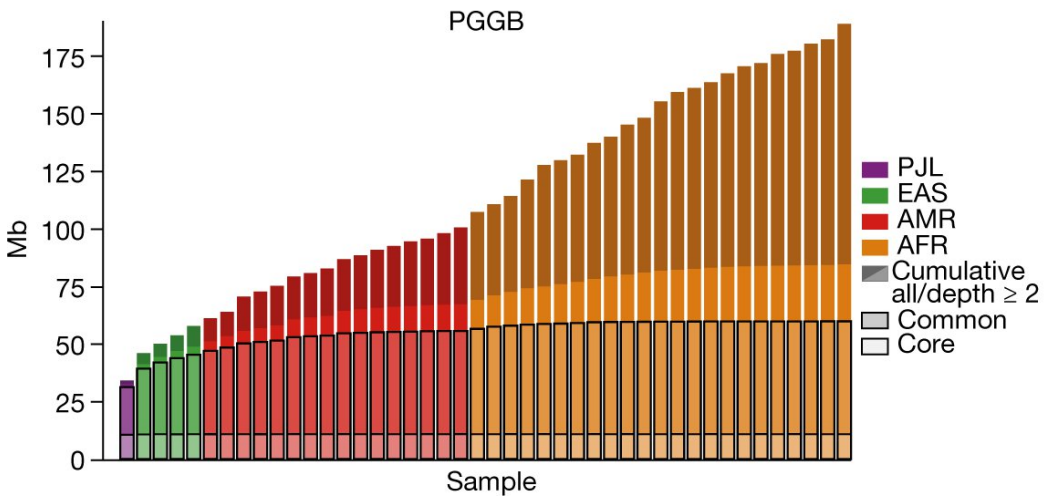
- Integrating every genomic variation from 10s to 100s of long genomes is not trivial !
  - Today 4 methods (from 2 research groups)
  - For 12 bovine assemblies (Leonard AS et al, 2023)
    - Graphs are huge !
    - Computational requirements can be huge !

Parameter	Unit	minigraph	cactus	pggb
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

- 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)



Andreace et al, 2023

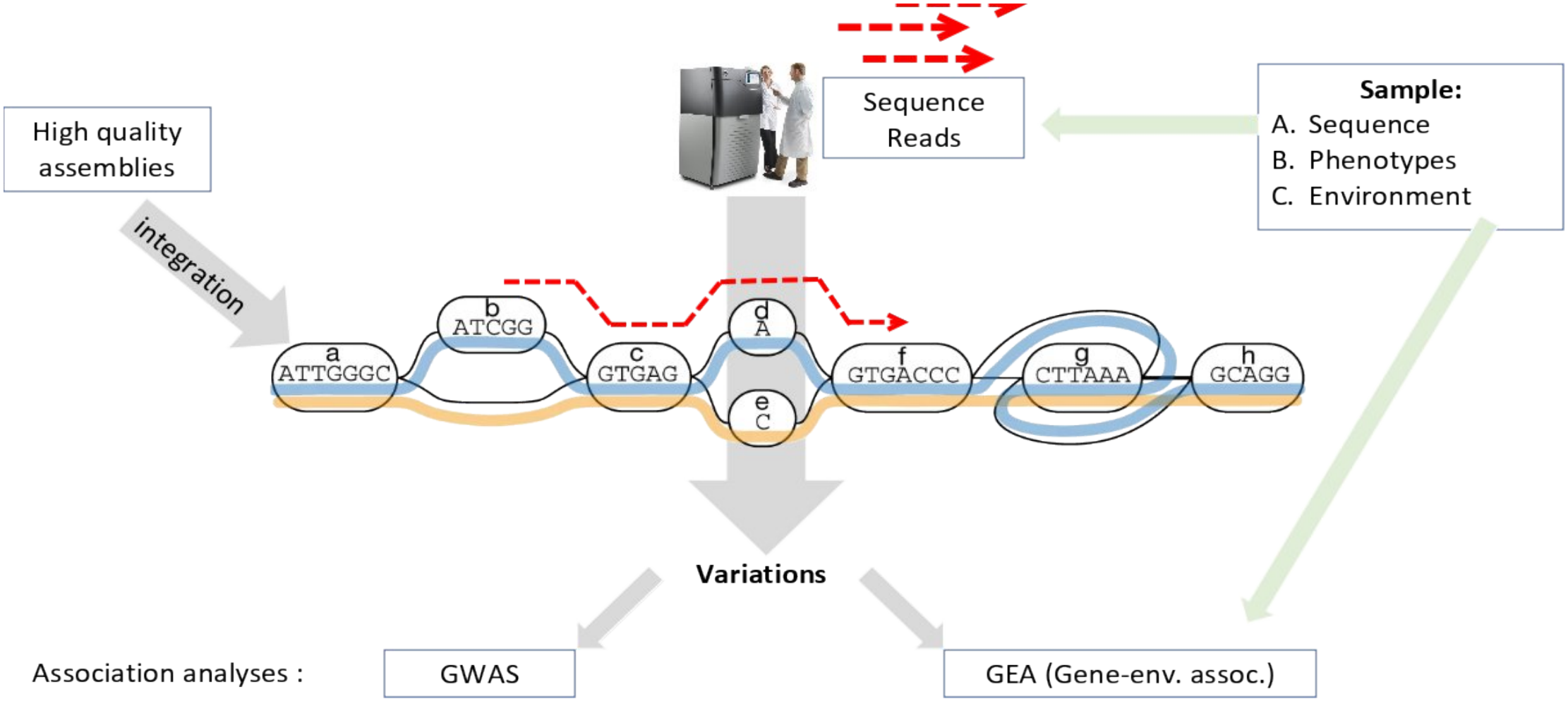
Tool	Metric	2 haplo.	10 haplo.	104 haplo.
MINIGRAPH	region			
	nodes	4	19	29
	edges	1	26	40
MINIGRAPH CACTUS	region			
	nodes	842	3543	--
	edges	1127	4016	--
PGGB	region			
	nodes	2	3824	--
	edges	0	5160	--

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 et appel de variants

- **Aujourd'hui : utilisation majoritaire = étude de variant et génotypage**



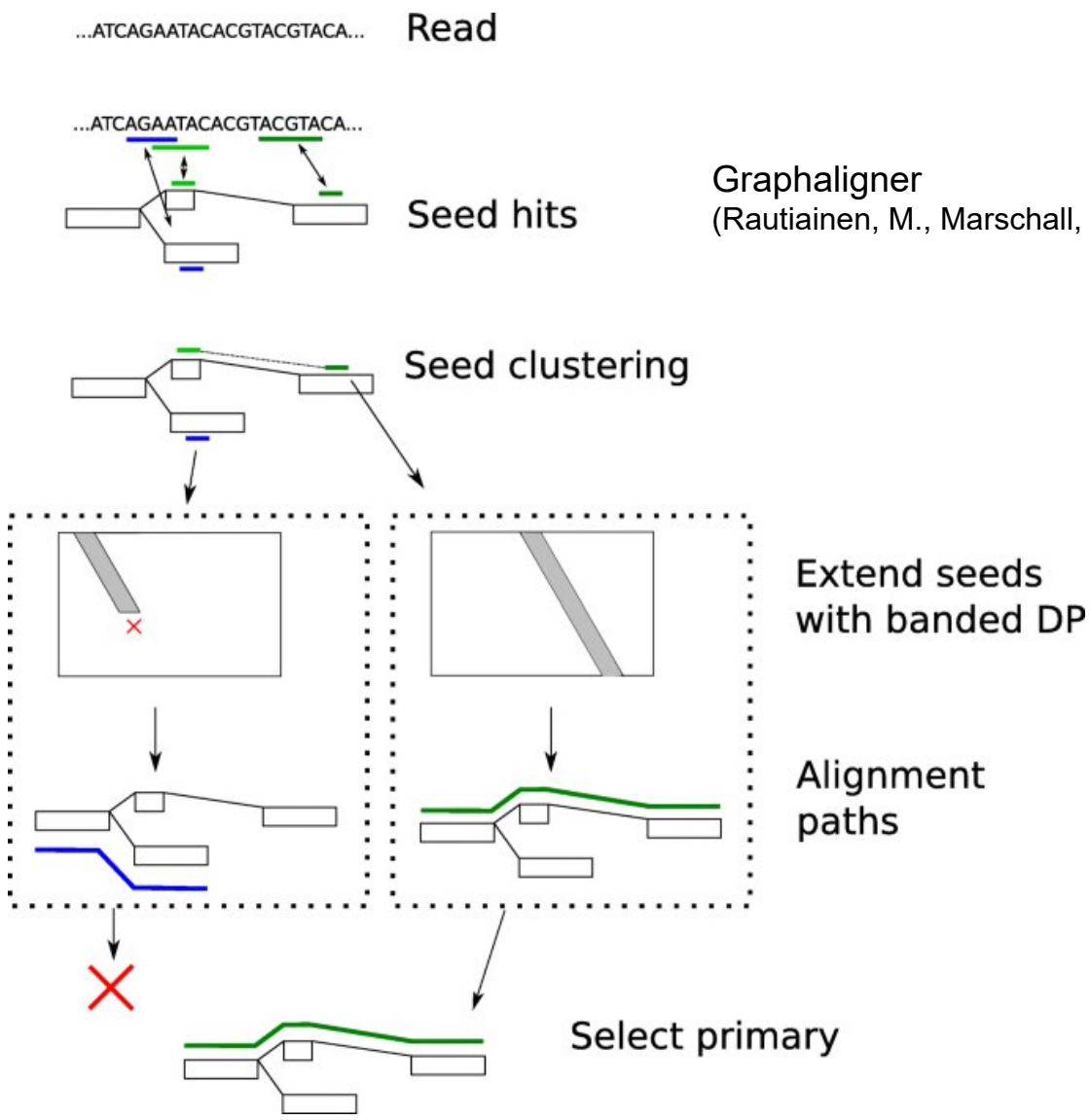
IV. Mapping et appel de variants

- “seed and extend” approach

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

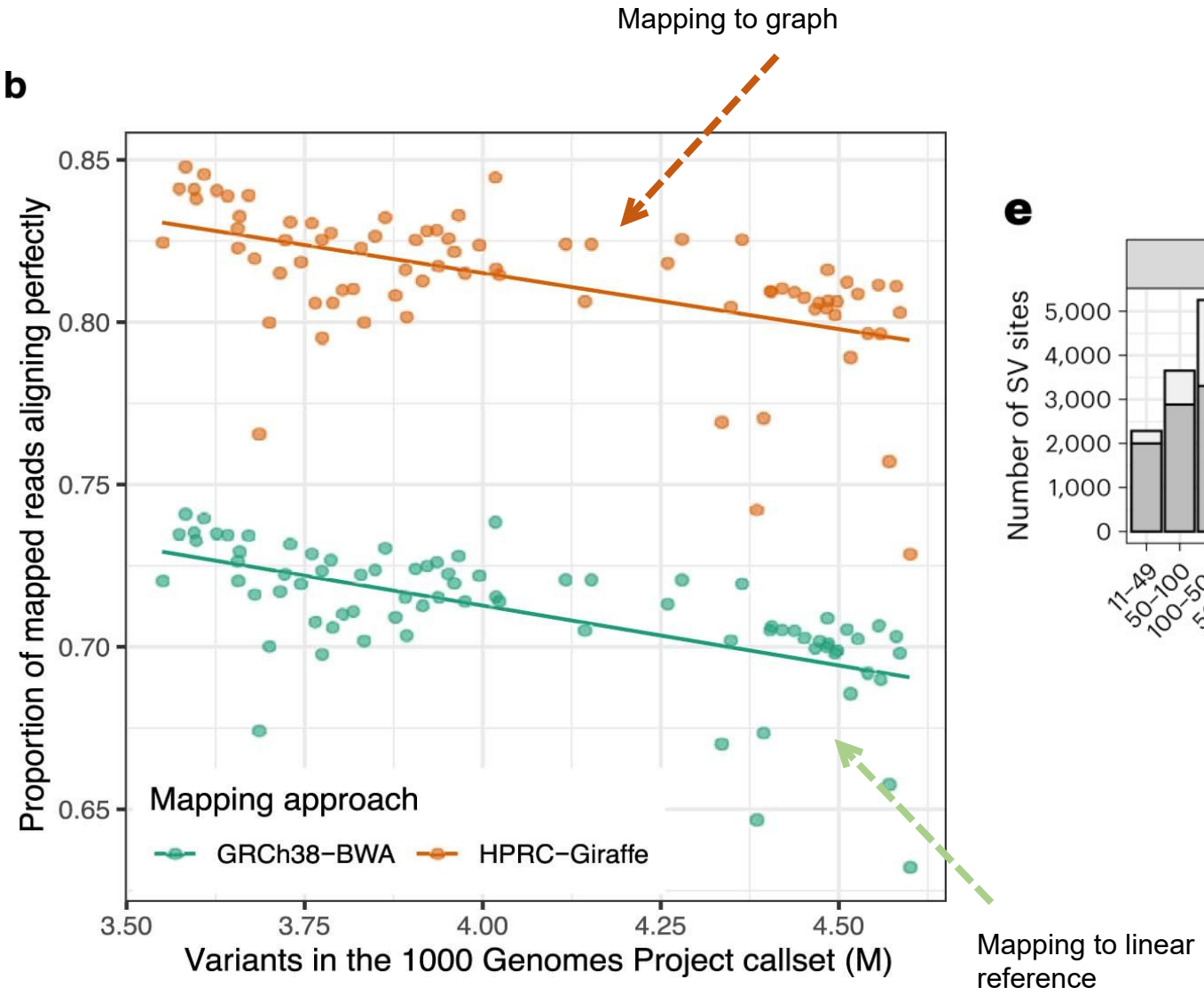
2. Mapping itself :

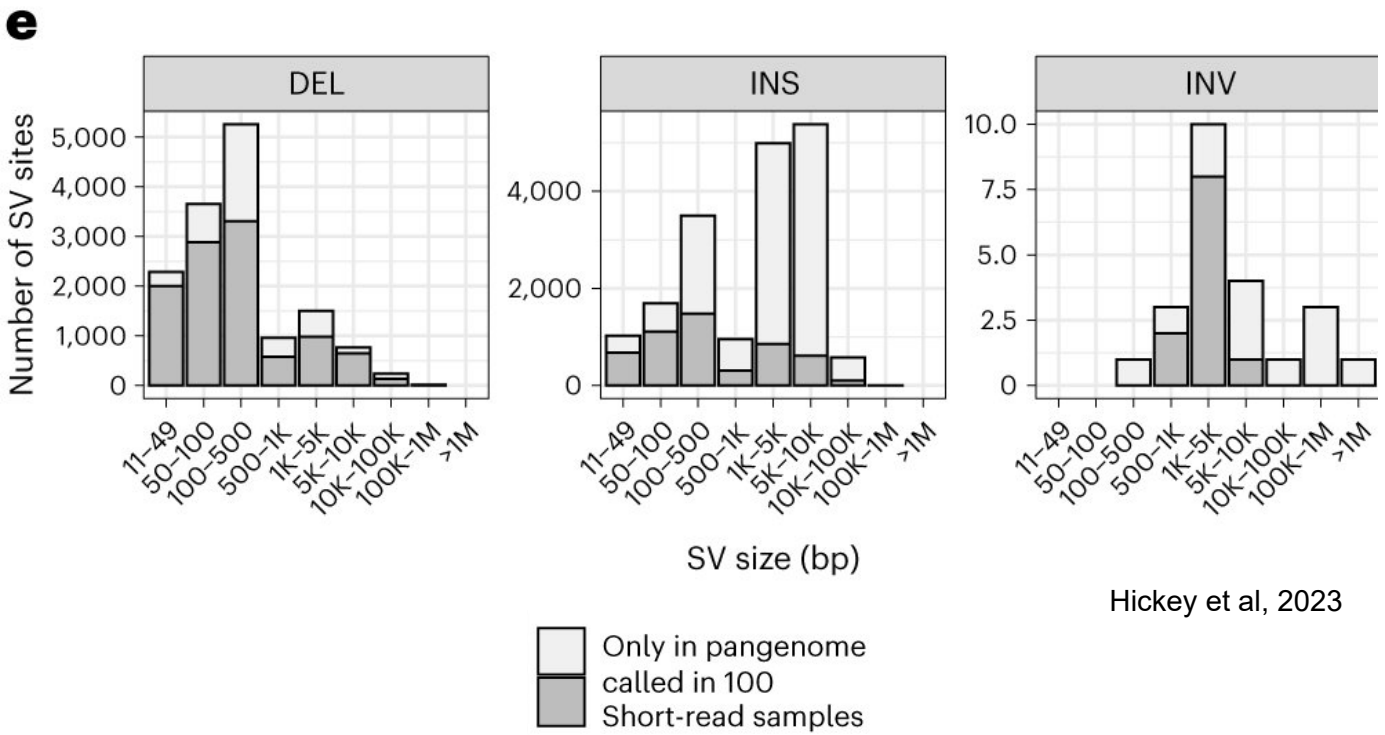
    1. Find seed anchors between query and graph nodes
    2. **From these anchor, select a path in the graph**
    3. Then “classic” alignment between query and sequence of selected path



Graphaligner
(Rautiainen, M., Marschall, T, 2020)

IV. Mapping et appel de variants





Hickey et al, 2023

Lia, WW. *et al.* (2023)

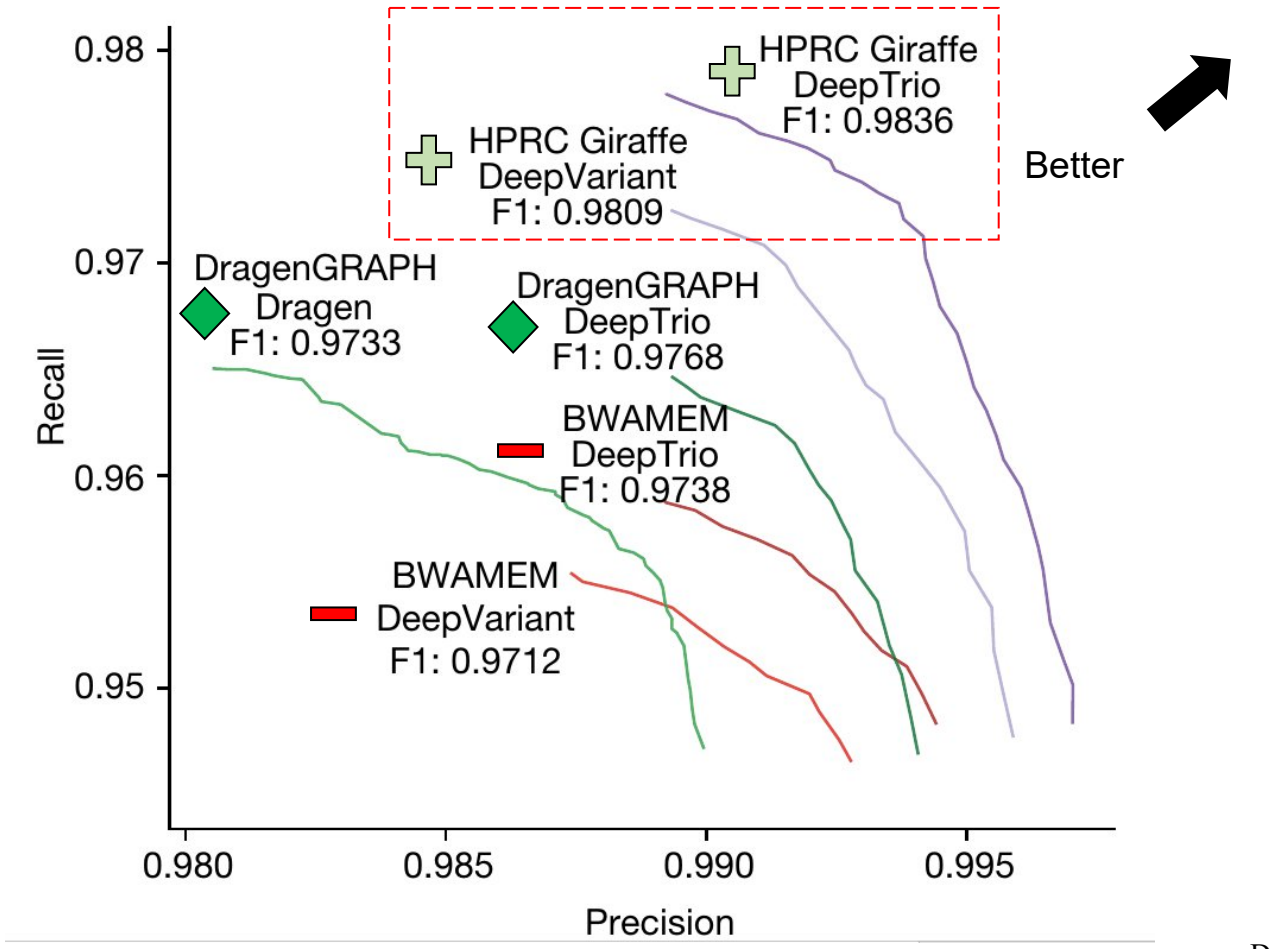
IV. Mapping et appel de variants

Genotyping performance gain for graph-aided analysis of short reads

Mapping to graph

Mapping to linear reference augmented with a variant database

Mapping to a single linear reference genome (standard approach)



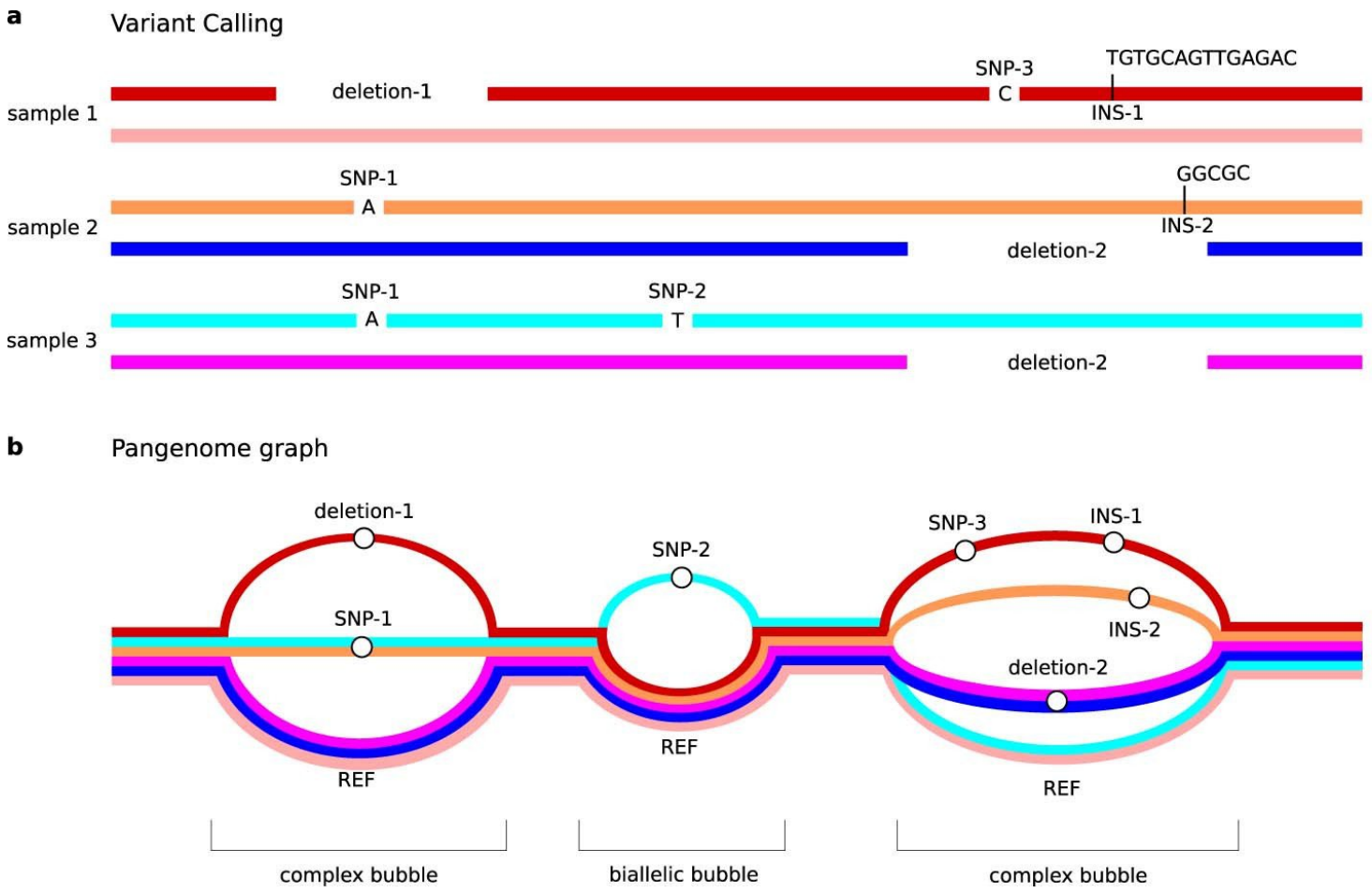
Liao et al, 2023

DeepVariant
DeepTrio
= deep learning-based variant callers

IV. Mapping et appel de variants

- Most methods rely on SNPs
- And reference-based variants (VCF)

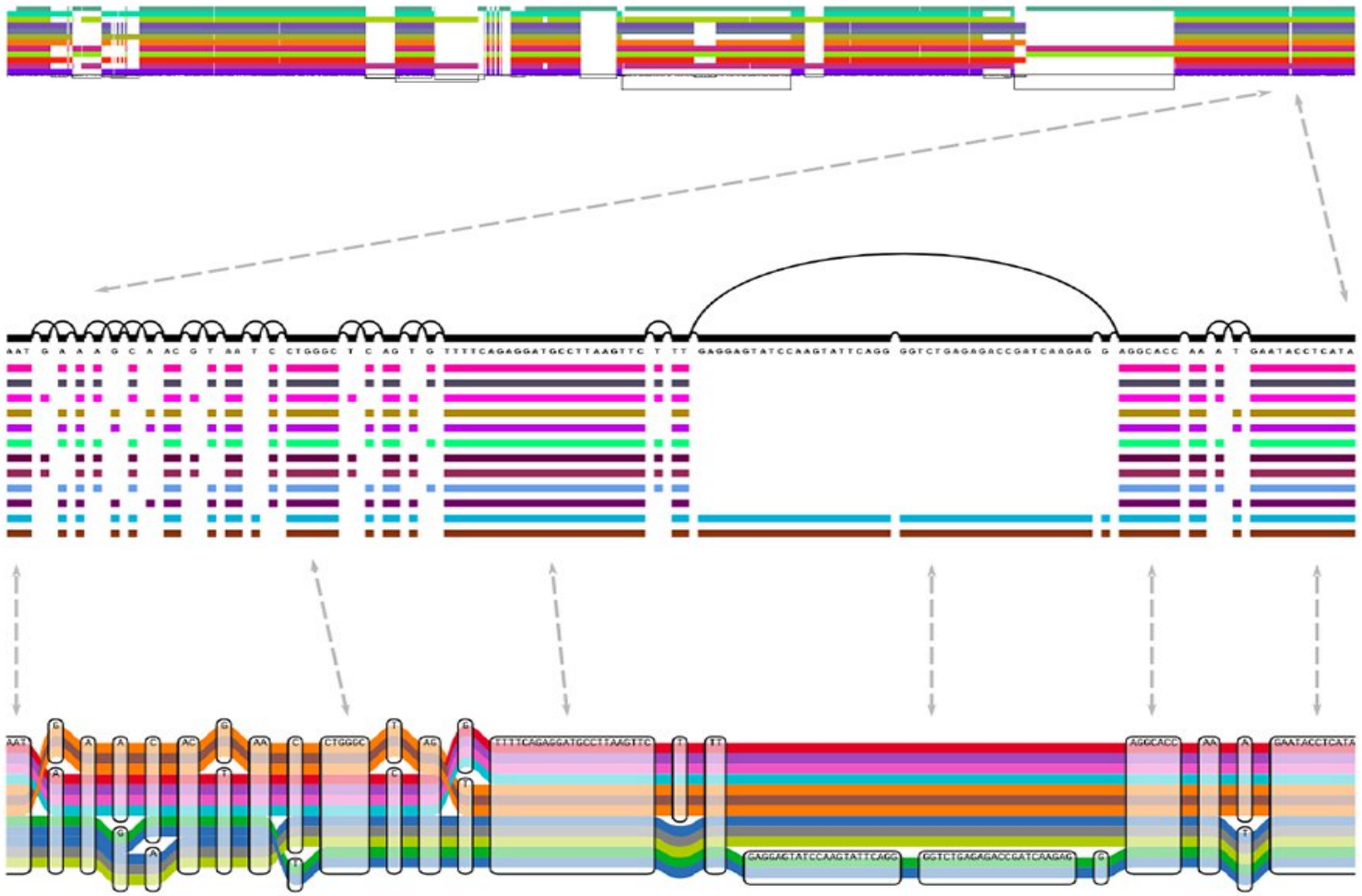
- More complex relationships can be extracted from the graph ! (ex: nested variants)



V. Visualisation

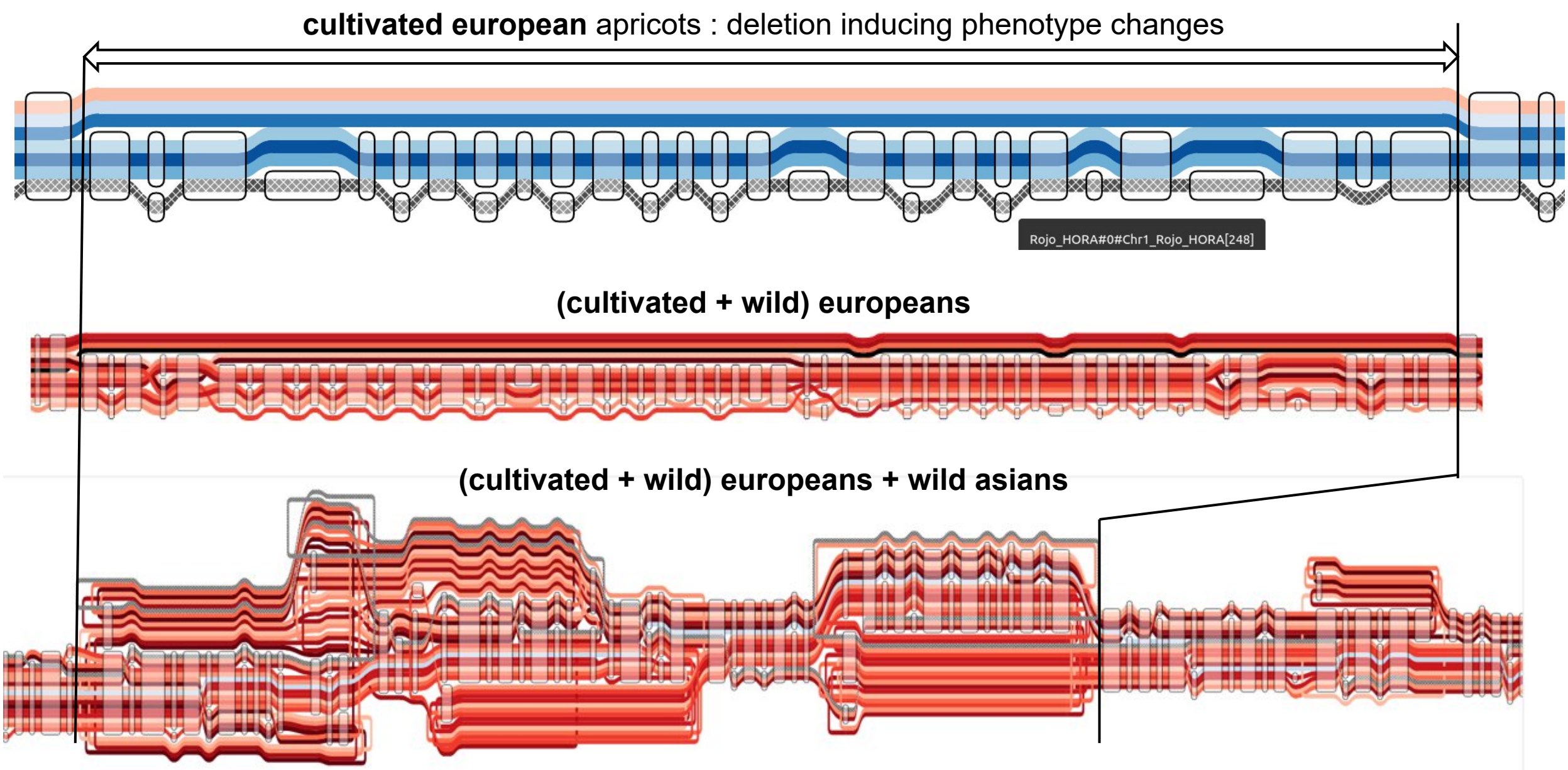
- Likely the most limited point today :
  - How to switch from coordinate-based to graph-based representation ?
  - How to make sense of huge & complex topologies in 2D ?

ODGI
(Guarracino et al, · 2022)



SequenceTubeMap
(unpublished)

V. Visualisation

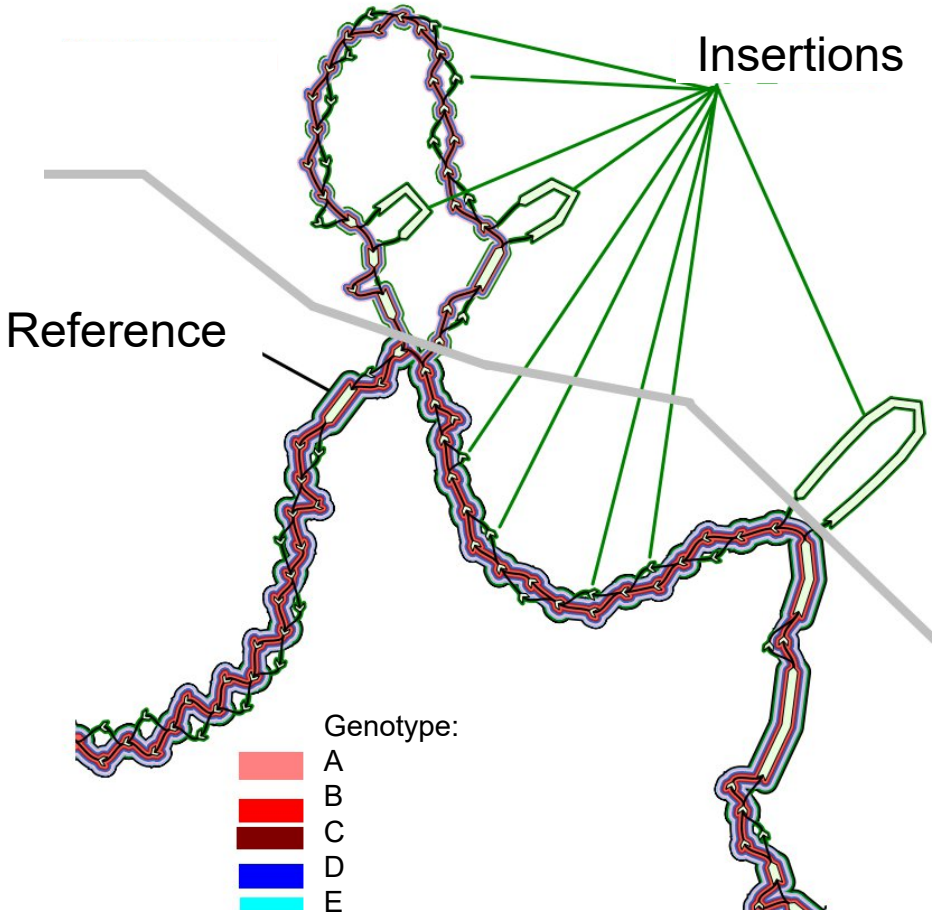


Courtesy of Ismael Blanchard, Véronique Decroocq.INRAE, BFP, Bordeaux

V. Visualisation

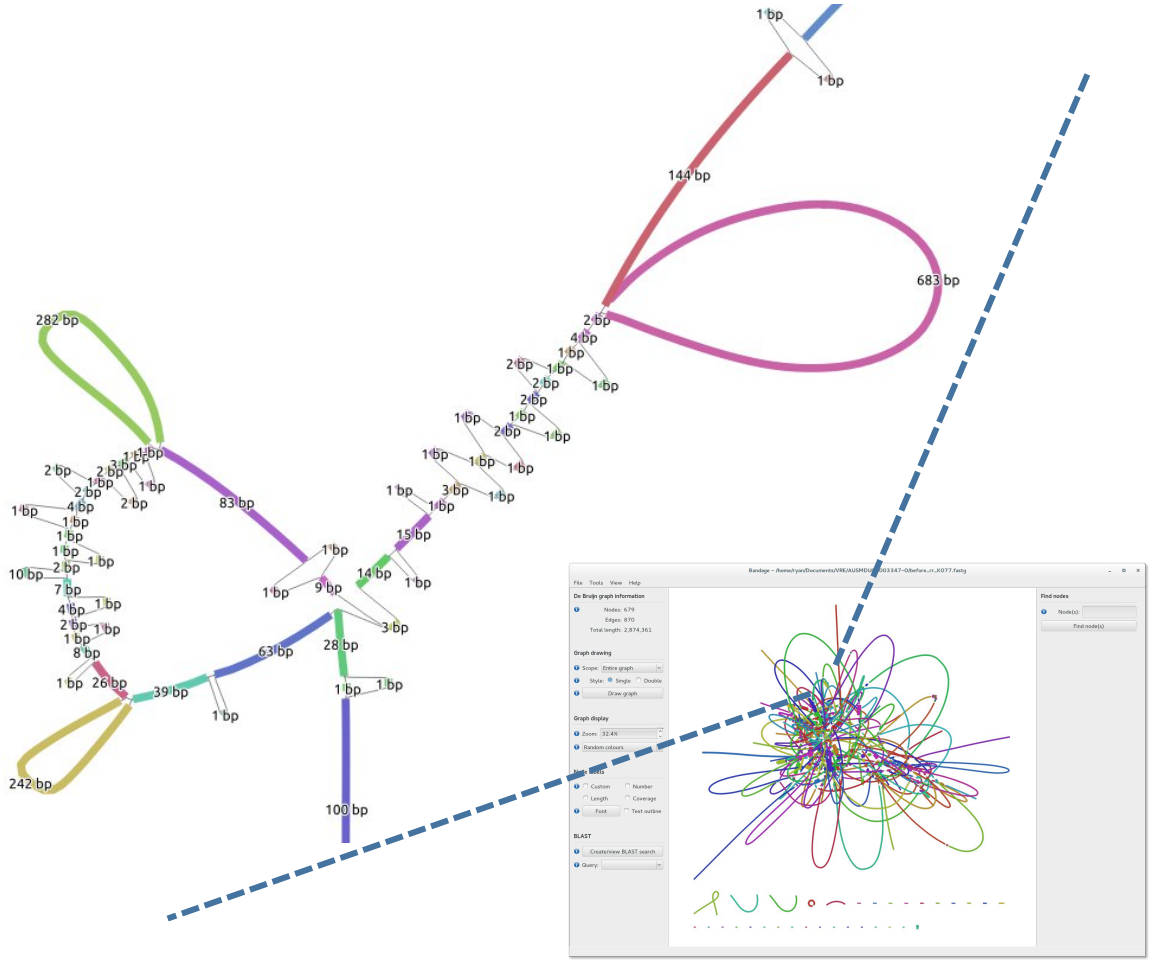
gfaviz (Gonnella et al, 2018)

+ customizable path annotations
+ adapted to figures generation
– limited to very small subgraphs



Bandage (Wick et al, 2015)

+ manipulation of very large graphs
– paths cannot be displayed



Thank you for your attention.

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