Simulating population pangenomes under coalescent demographic models

lucienpiat33@gmail.com

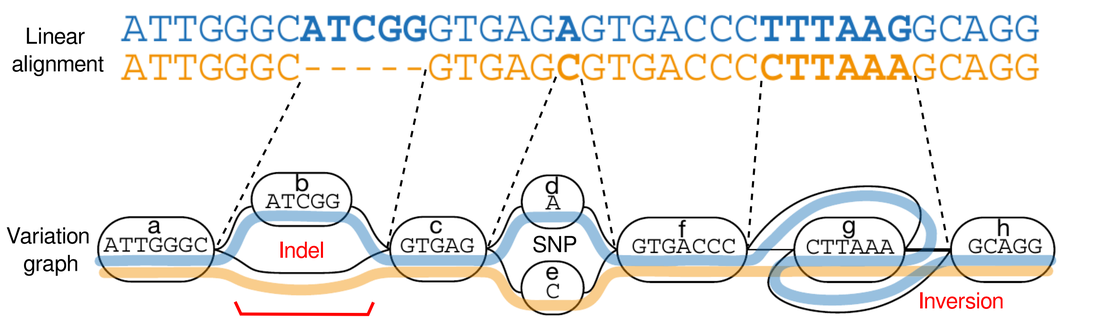
Lucien PIAT[1], Sukanya DENNI[1], Siegfried DUBOIS[2,5], Florent COUTURIER[1], Nicolas LAPALU[3], Benjamin LINARD[4], Claire LEMAITRE[5] & Ludovic DUVAUX[1]

1.INRAE UMR Biogeco 1202, Domaine de l'hermitage 69 route d'Arcachon, 33612, CESTAS, France
2.INRAE UMR GenPhySe Cytogene 1388, 24 chemin de Borde-Rouge Auzeville Tolosane, 31326, Castanet Tolosan, France
3. INRAE UR BIOGER, 22 place de l'agronomie, 91120, Palaiseau, France
4.MIAT INRAE Occitanie-Toulouse, Auzeville, 31326, Castanet Tolosan, France
5.Inria IRISA UMR 6074, Campus de Beaulieu, 35042, Rennes, France

Corresponding author :
ludovic.duvaux@inrae.fr

Structural variants and pangenomes

Context: **Pangenome** variation graphs gain more and more interest to comprehensively represent genetic diversity including **structural variants**.



Adapted from Liao et al. (2022)

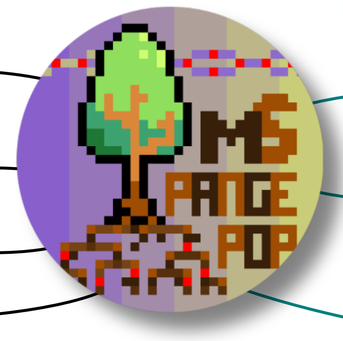
Structural variants (SVs): >50bp variations that change chromosome structure.

Problem : we don't know how to simulate realistic pangenomes

We present MSpangepop, a new tool for this purpose :

Features :

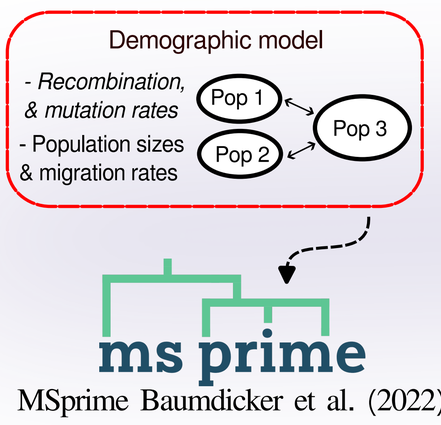
- Consider complex **demographic scenarios**
- Realistic (nested) **variants**
- Produce large genomes
- Export graph to GFA



Use cases :

- Tool benchmarking
- Genome evolutionary inferences
- OOP library to insert SVs into graph

Algorithm

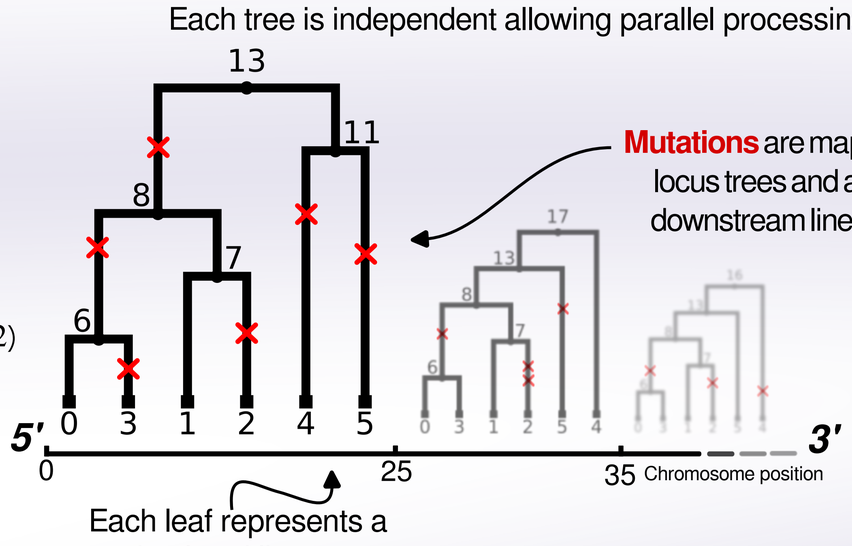


MSprime Baumdicker et al. (2022)

Produces genome genealogies

A. Ancestry simulation

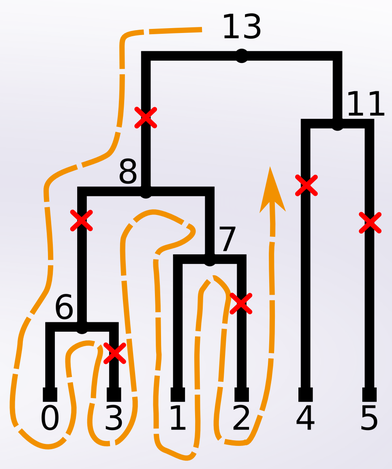
Along chromosomes, each recombination block hold a genealogical tree. Each tree is independent allowing parallel processing.



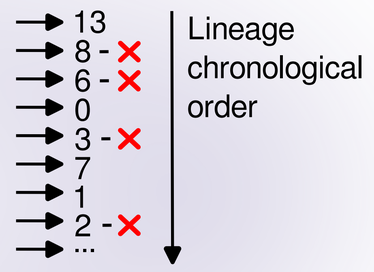
Mutations are mapped to locus trees and affect downstream lineages

Each leaf represents a simulated lineage

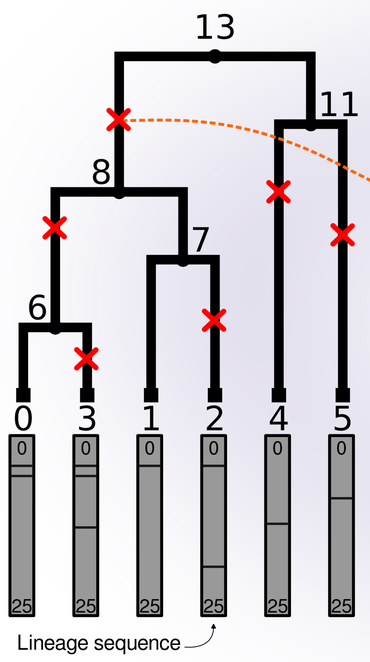
B. Recording mutations chronologically



Lineages are processed in turn and **mutations** recorded chronologically via **pre-order traversal**.

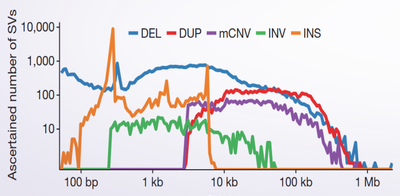


C. Adding structural variants with the MSpangepop library

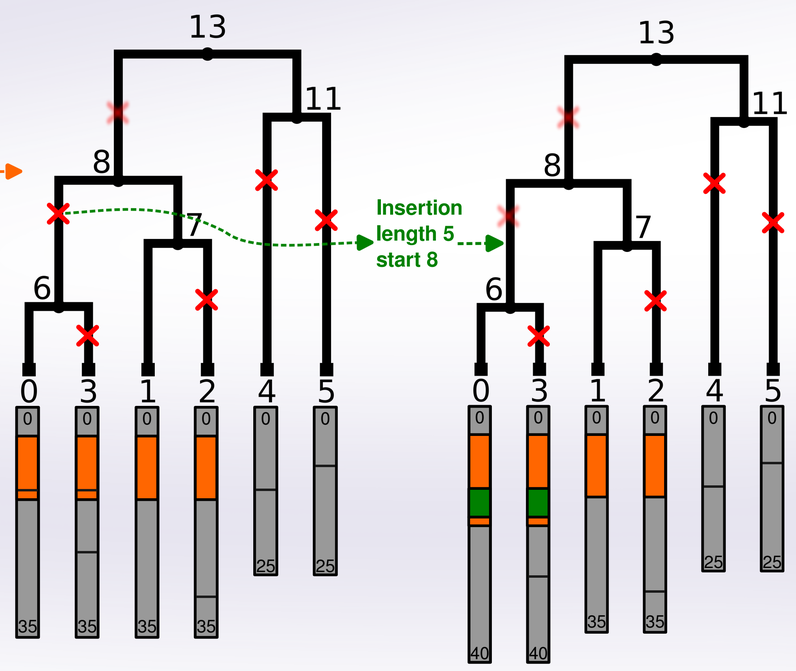


Variant type (SV or SNP) and length is drawn from a specified distribution

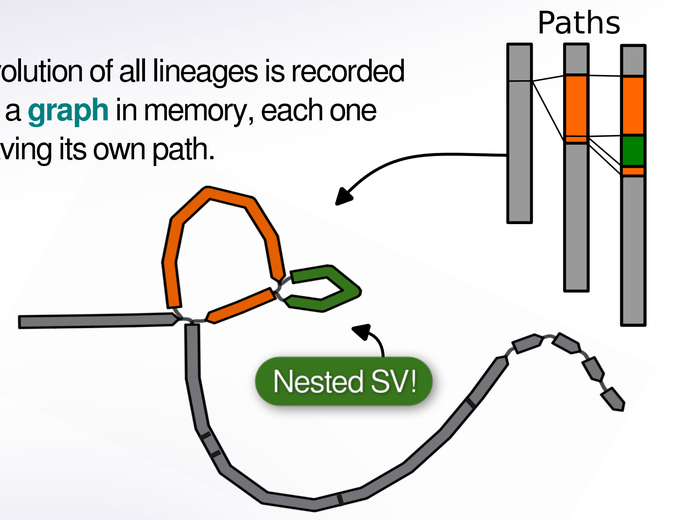
Insertion length 10 start 5



Sudmant, et al. (2015)

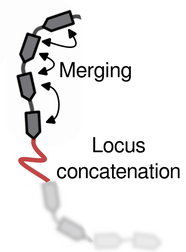


Evolution of all lineages is recorded as a **graph** in memory, each one having its own path.

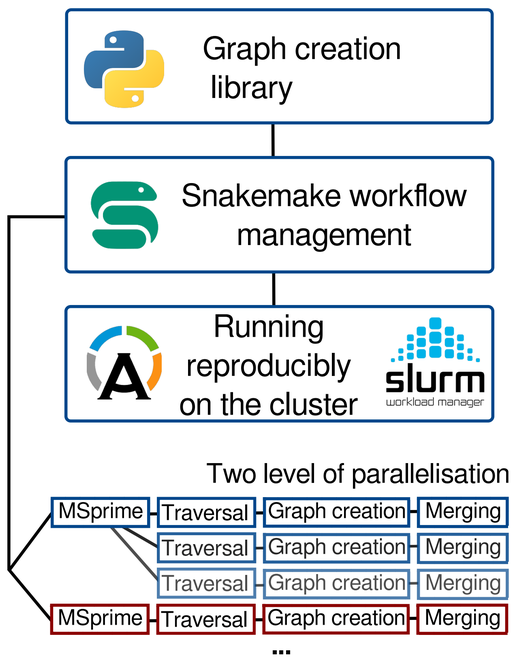


D. Node merging and locus concatenation

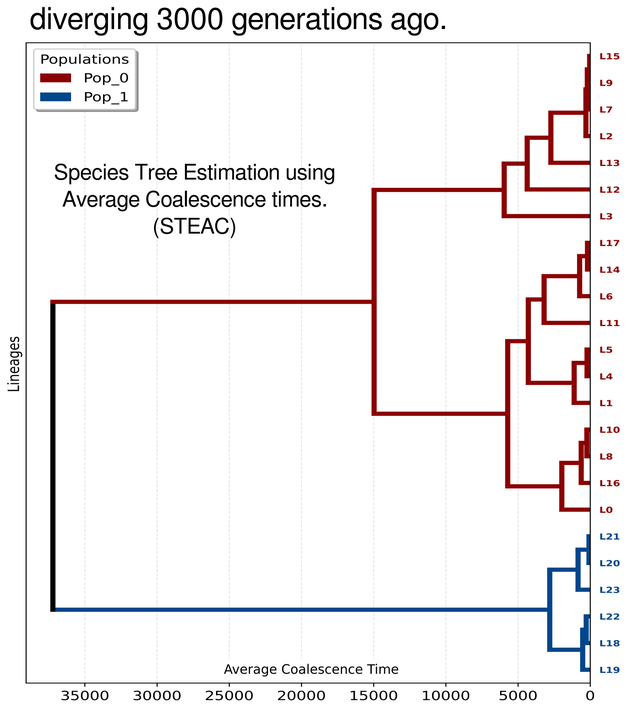
Output to GFA



MSpangepop overview

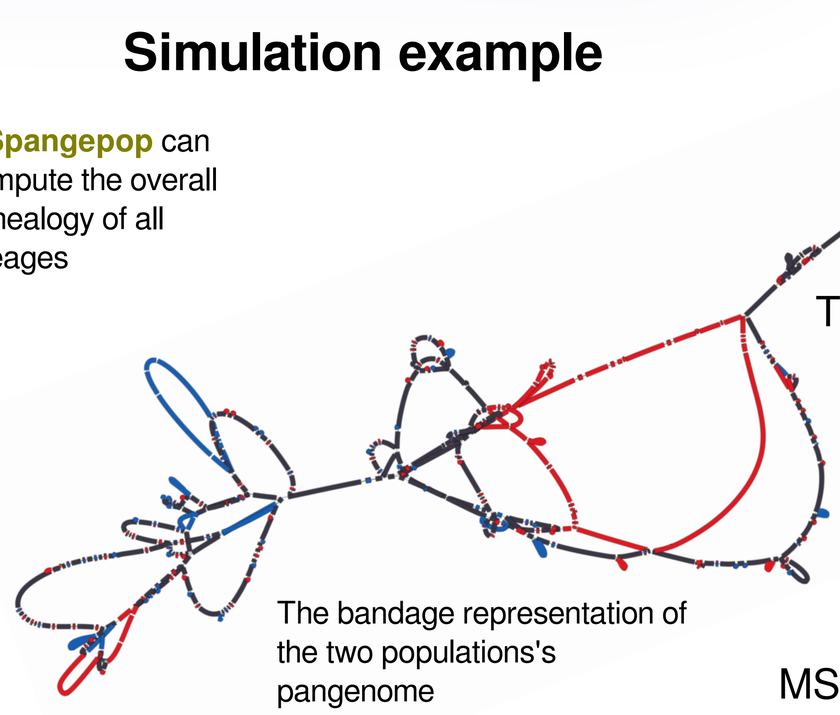


MSpangepop simulation of a *Staphylococcus aureus* **pangenome**, showing two groups diverging 3000 generations ago.



Simulation example

MSpangepop can compute the overall genealogy of all lineages



The bandage representation of the two populations's pangenome

This poster as PDF

MSpangepop repo

INRAE · université de BORDEAUX · Inria · MIA TOULOUSE · biogeco · 25 ans Master Bioinformatique · Normandie Université · UNIVERSITÉ DE ROUEN NORMANDIE · UFR Sciences et Techniques · CFC CFA CENTRE UNIVERSITAIRE DE FORMATION CONTINUE ET PAR ALTERNANCE