A Look at Trails through the Pangenome Visualization Jungle

Éloi Durant[1,2,3,4], François Sabot[1,4], Matthieu Conte[2] and Mathieu Rouard[3,4]

[1]DIADE, Univ Montpellier, IRD, CIRAD, Montpellier, France; [2]Syngenta Seeds SAS, Saint-Sauveur, France; [3]Bioversity International, Parc Scientifique Agropolis II, Montpellier, France; [4]French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier, France

Pangenomes are inventories of genomic material found in related genomes. Applied first to bacteria, they were defined as sets of genes that might be shared between strains, containing a **core** genome (genes present in every strain) as opposed to the **dispensable** genome (Tettelin et al., 2005). To fit eukaryotic organisms this definition later evolved, now including all chunks of sequences and their succession. Their representations evolved accordingly, from sets to graph but without any user-friendly tool for their effective visualization applied to eukaryotes. Our efforts toward a user-friendly visualization tool led to the representation of **linearized** pangenomes.
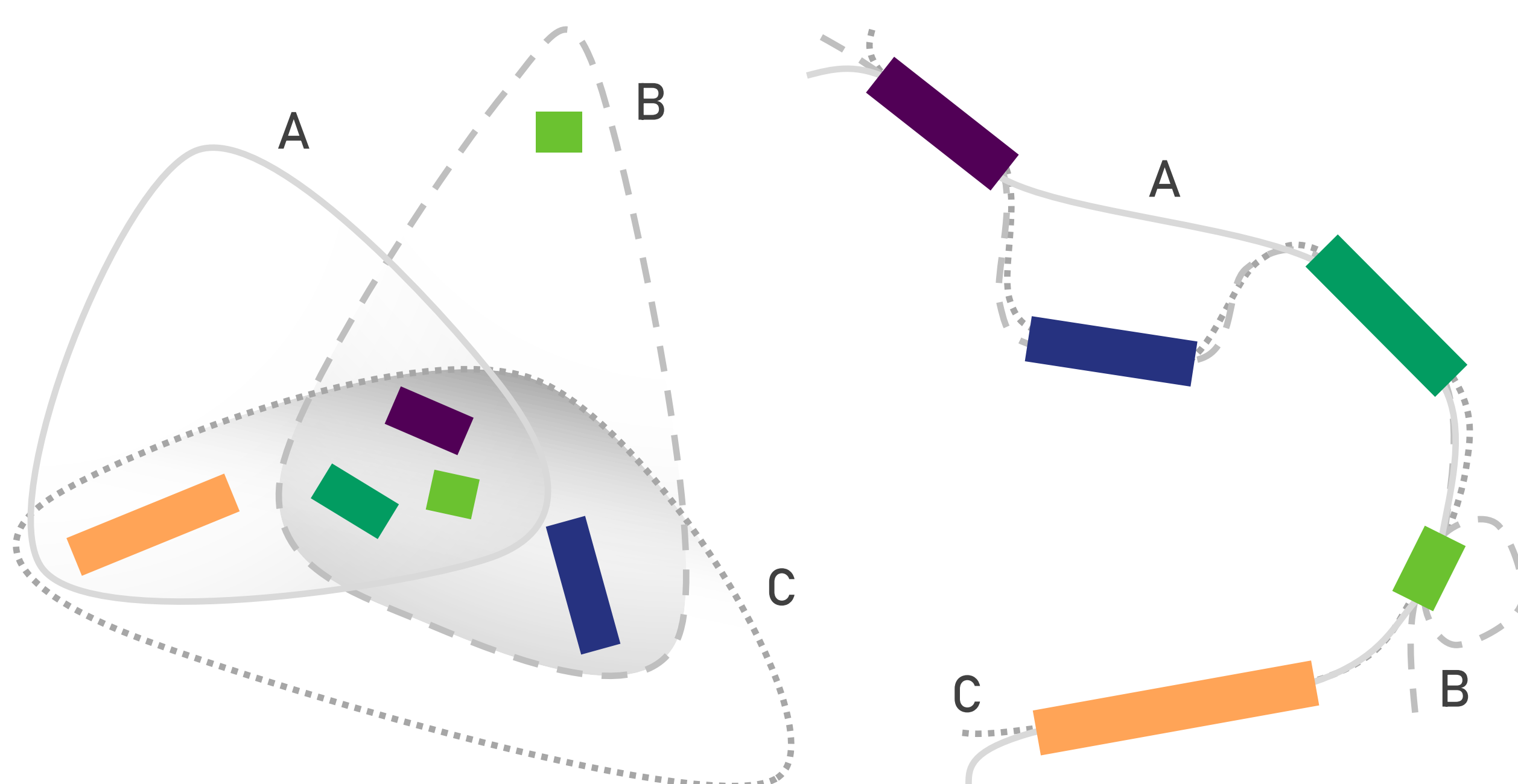
Venn diagrams are usual representations for sets. They lack information of position, and do not scale well with the number of sets. Alternatives like UpSet diagrams face similar scaling problems with dozens of genomes / sets.

Sequence graphs are recent representations, with every genome being a **path** through the sequence nodes of the graph. New genomes create branches, ending in visual clutter. They can be visualized with tools like Bandage or GfaViz.

Using a **linear coordinate system** can improve the readability and explorability. Existing implementations include UCSC's **snake tracks** and **PGAP-X**. A presence / absence matrix can be added to infer succession in the genomes.

Linearity, while common for the visualization of genomes and their alignments, is not widely used for pangenomics. Closest examples include visualizations of **positioned genes** on a linear coordinate system—without information on the intergenic space—or visualizations displaying all **structural variations** from the genomes—with the drawback of being hard to read and space consuming.

We therefore introduce Panache, our PANgenome Analyzer with CHromosomal Exploration, a browser-based interface created for the exploration of linear representations of pangenomes. It displays *pangenomic blocks* (either genes or sequences) as ordered on a single string in a genome browser-like fashion, with a set of tracks displaying summary information as illustrated on the right → → → → → →

A demo version with datasets and documentation is already available on GitHub!

Tettelin, H., et al., *Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"*. Proceedings of the National Academy of Sciences of the United States of America, 2005.

Core threshold
- Custom categorization of your *core* and *dispensable* genomes

Minimal Presence ratio to be part of *Core*

85%

Files

bananache_chromOnly.tsv — Browse

Musaceae-pangenome.gene.gff — Browse

Chromosome on display

chr01

Display parameters

Shape type

Minimal Presence ratio to be part of Core

85%

Zoom Level

Hollow area finder

Absence rate (0-1) — 0,4

Hollow area finder

Absence rate (0-1) — 0,4

Number of consecutive blocks — 2

There are **15** regions matching these criteria

1 Before	14 After

117519 Targeted position

-114383 | ◀ ▶ | +812330

Annotation cards
- Displays annotations from gff on hovering

Ma01_g00080
Gene's gff START: 61858

62k 63k 64k 65k 66k 67k 68k

Note: serine/threonine-protein phosphatase PP-X isozyme 2-like

Miniature
- Overview of full panchromosome
- Navigation on click

PAV matrix
- Shows present and absent blocks (columns) for each genome (rows)

0 2M 4M 6M

0 2M 4M 6M 8M 10M 12M 14M 16M 18M 20M 22M 24M 28M

Bile
BSK30
Calcutta
Ensete01
EnseteBedadit
EnseteDerea
FHIA
Itinerans
Kole
Lidi
MasKirana
Pahang

chr01
chr02
chr03
chr04

This block appears in 11 selected genome(s)

Hollow Area Finder
- Quickly jumps to areas of consecutive absence

Sorting Options

NEW

Sort the tracks

None
Alphanumeric
Reverse alphanumeric
Local presence/absence pattern
Gene presence status
Phylogenetic tree

Info tracks
- Sums up information per block
- Can show distribution of repeated blocks