13. Introduction to pangenomic

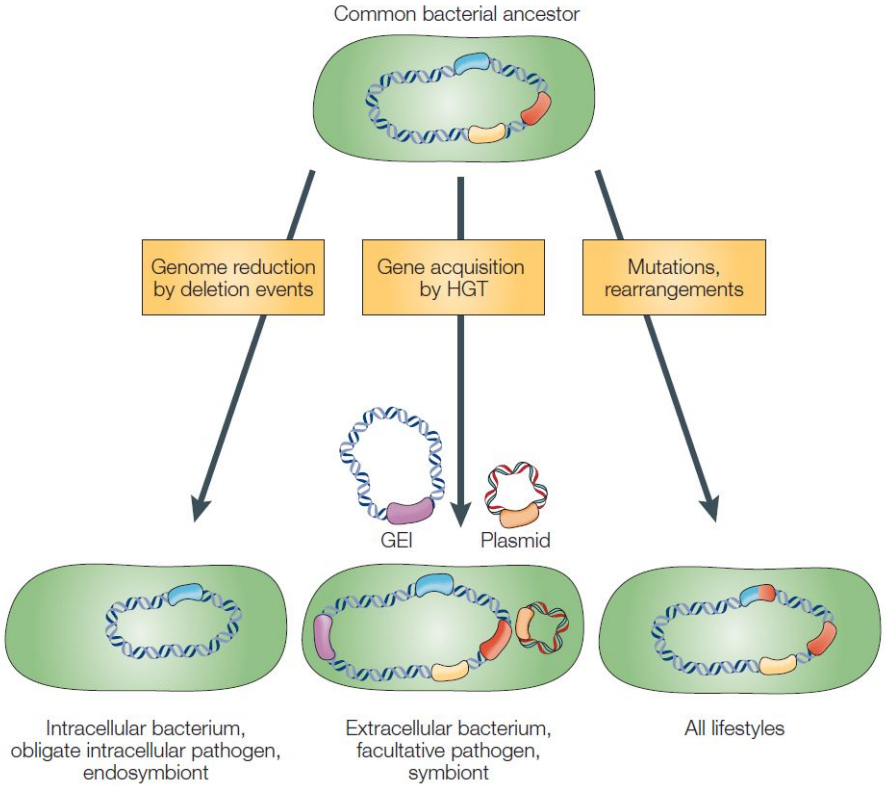
Guillaume GAUTREAU, 29/09/2022

A bit of oriented and simplified history

- 1928: Griffith experiment, DNA is the molecular support of life information
- 1953: discovery of the DNA structure of DNA
- 1977: Sanger sequencing (phi X 174 is sequenced)
- 1985: Invention of PCR
- 1977: Woese's tree of life base on 16s RNA
- 1995: *Haemophilus influenzae*, first bacterial genome is sequence
- 2000s: hightroughput sequencing : 454, Solexa
- 2000s: plenty of bacterial genomes for each species
- from 2005 up to now:
    - Surprise: **there is a lot a diversity in the genomes of a same species**

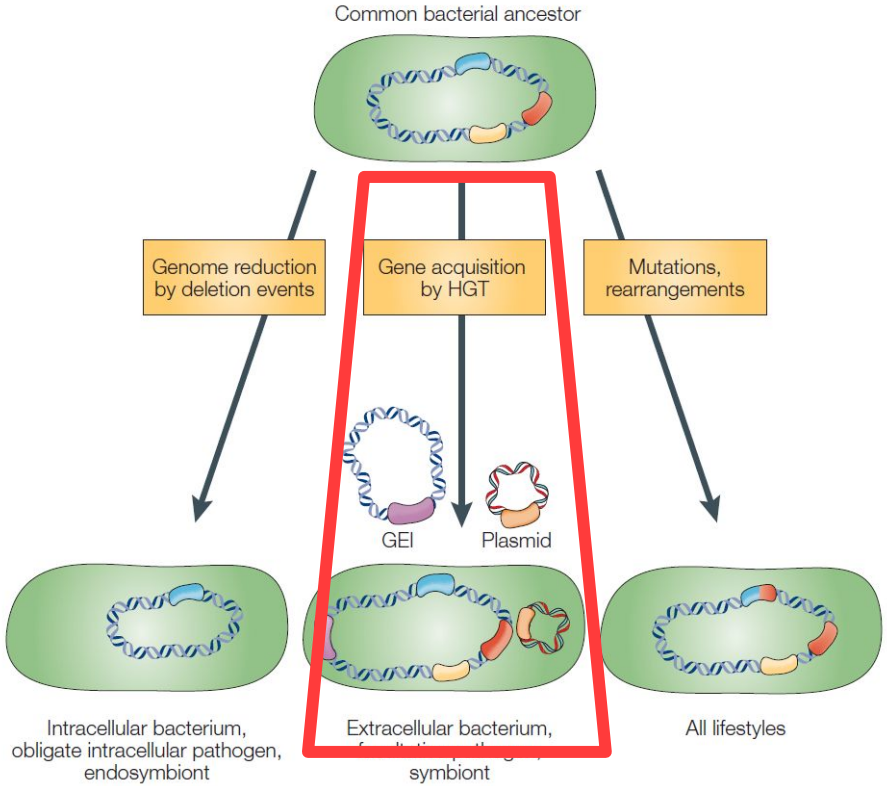
Procaryote evolution paths

Figure from Dobrindt et al., 2004



Procaryote evolution paths

Figure from Dobrindt et al., 2004

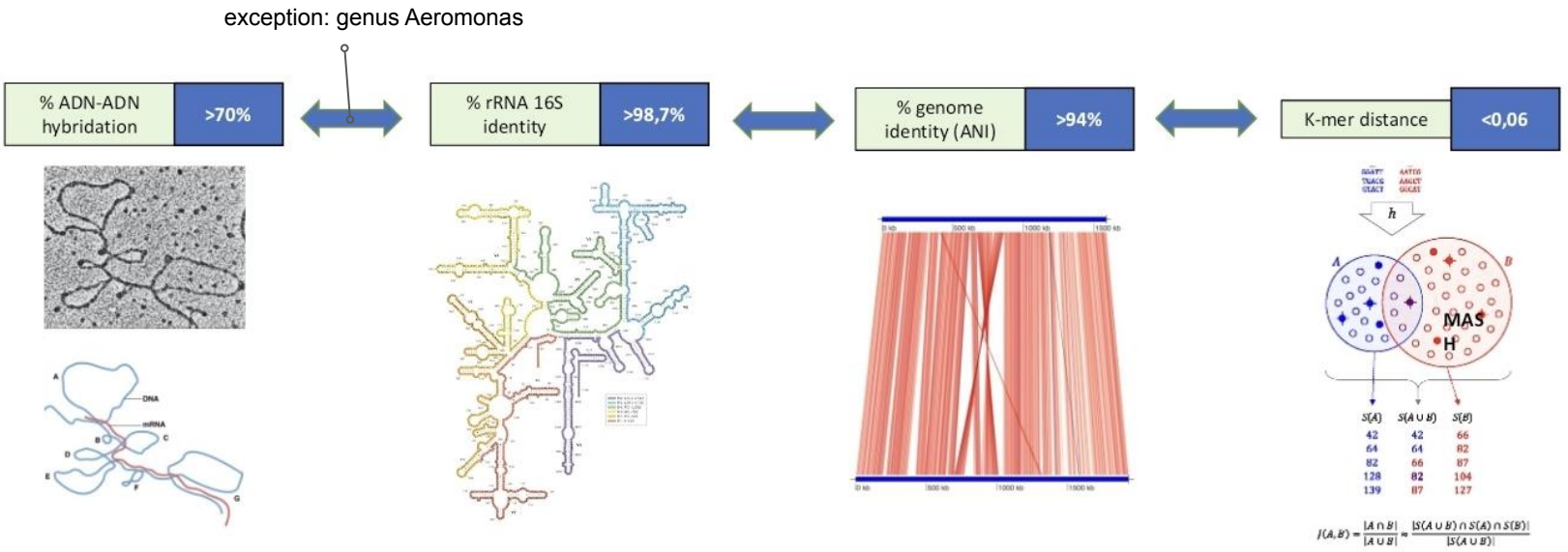


(Treagen et Rocha, 2011)

Horizontal Gene Transfers (HGT), not duplication, drive the expansion of protein families in procaryotes

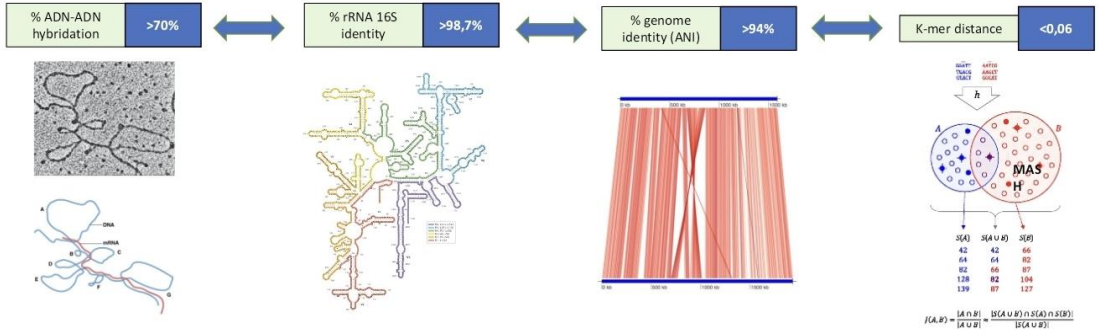
Diversity in numbers

- Reminder: Species can be defined as ”genomically coherent group of organisms”:



Diversity in numbers

- Reminder: species can be defined as ”genomically coherent group of organisms” :

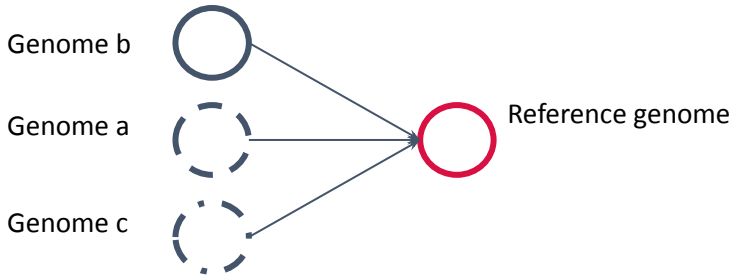


⇒So: high level of polymorphism (compare to human genomes for example)

- High diversity of gene content from HGT (5% to 30-40% of variable genes)
  - Intracellular: low diversity (mostly clonal)
  - Free living bacteria: high level of diversity to fit many ecological niches

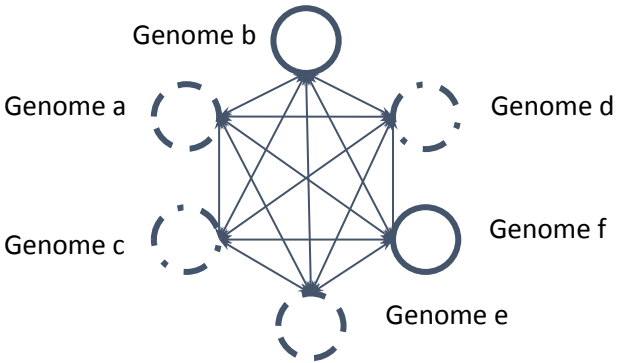
How to analysis all of this diversity ?

Known reference



⇒ **Exhaustivity issue**

No reference



⇒ **Combinatorial problems for interpretation**

The concept of pangenome (also named pangenome)

The term pangenome was first coined in 2005 (Tettelin et al., 2005) and describes:

- The union of sequence entities shared by genomes of interest (usually genomes of a species)

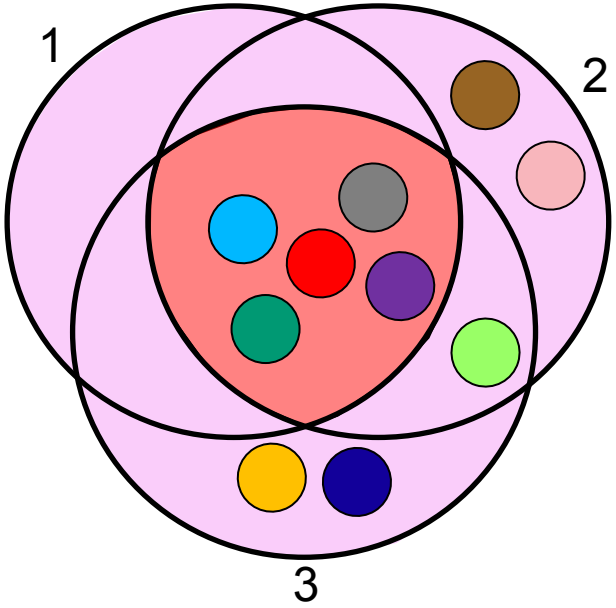
At the crossroads of comparative genomics, population genomics, phylogenomics and metagenomics, pangenomics offers to:

- Compile and organize the genomic diversity trough compact data structures
- Shed light on common elements but also variable ones

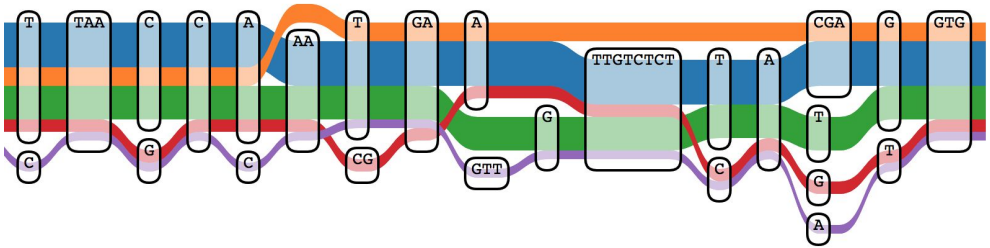
Suitable for bacteria because hundreds of genomes are now available for many species but also relevant for human and plant studies (and many more…)

Two main levels of understanding for pangenomic

1. Gene families based pangenome

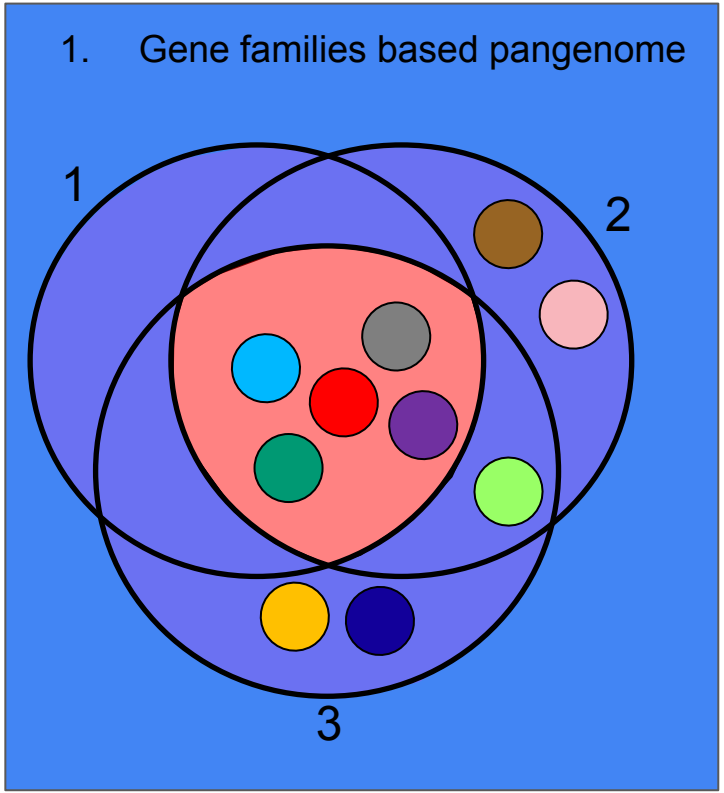


2. Sequence based pangenome (often via a sequence graph)

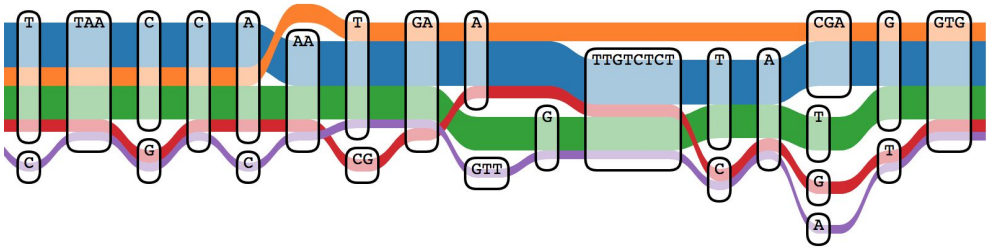


Two main levels of understanding for pangenomic

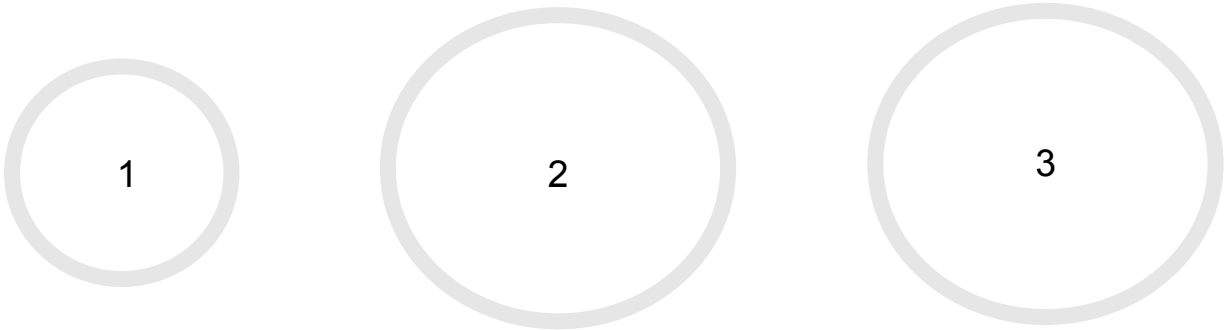
1. Gene families based pangenome



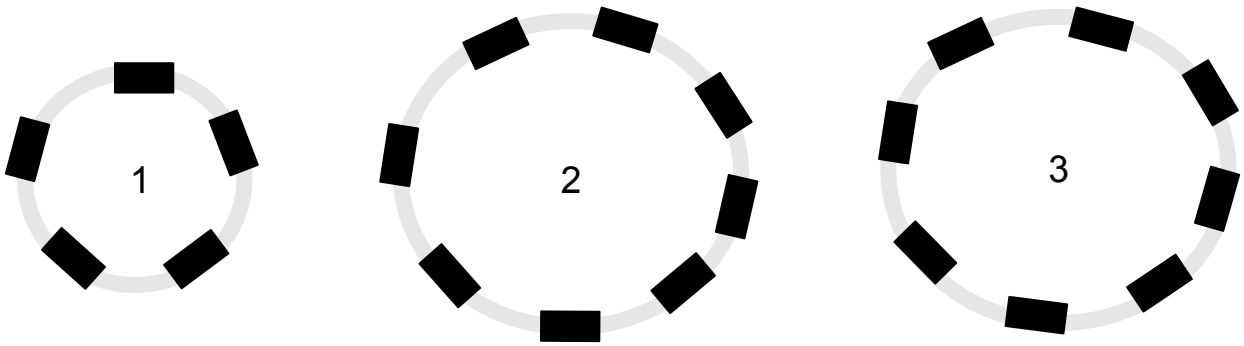
2. Sequence based pangenome (often via a sequence graph)



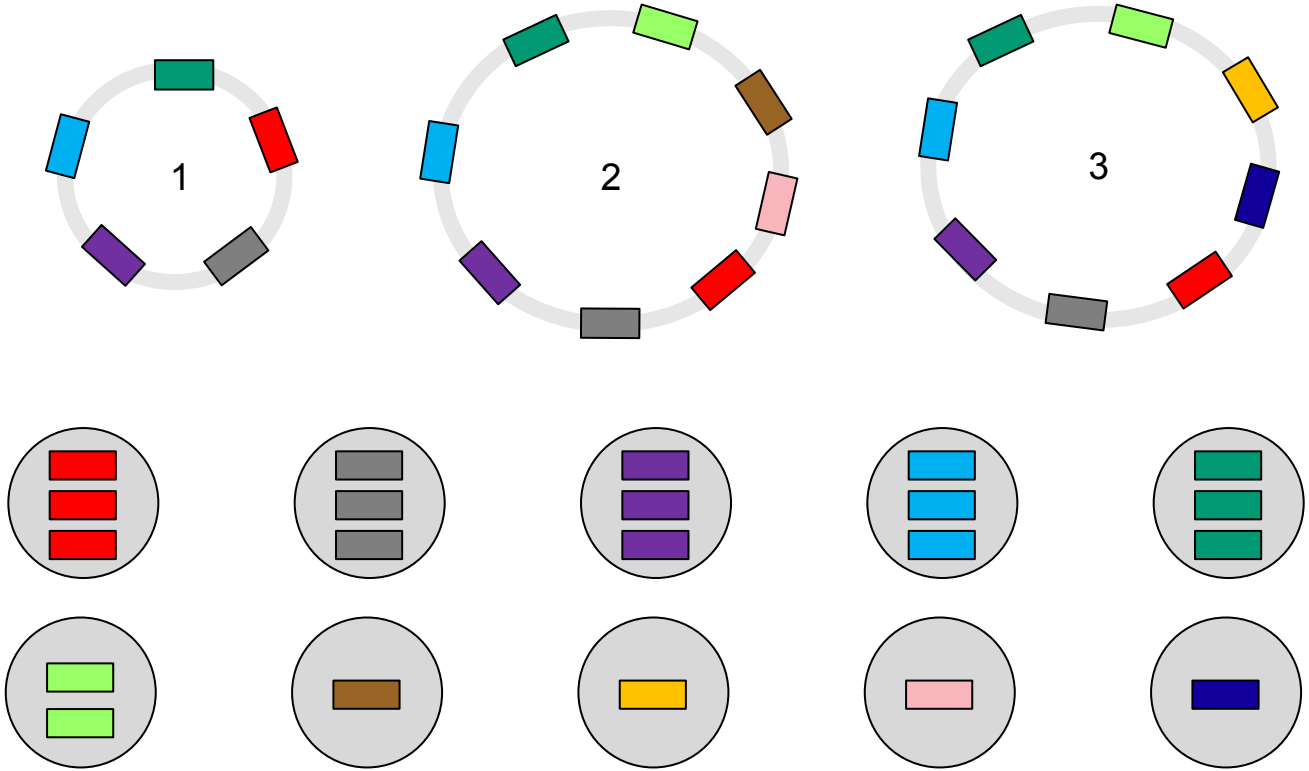
Toy example about a way to create prokaryotic pangenome



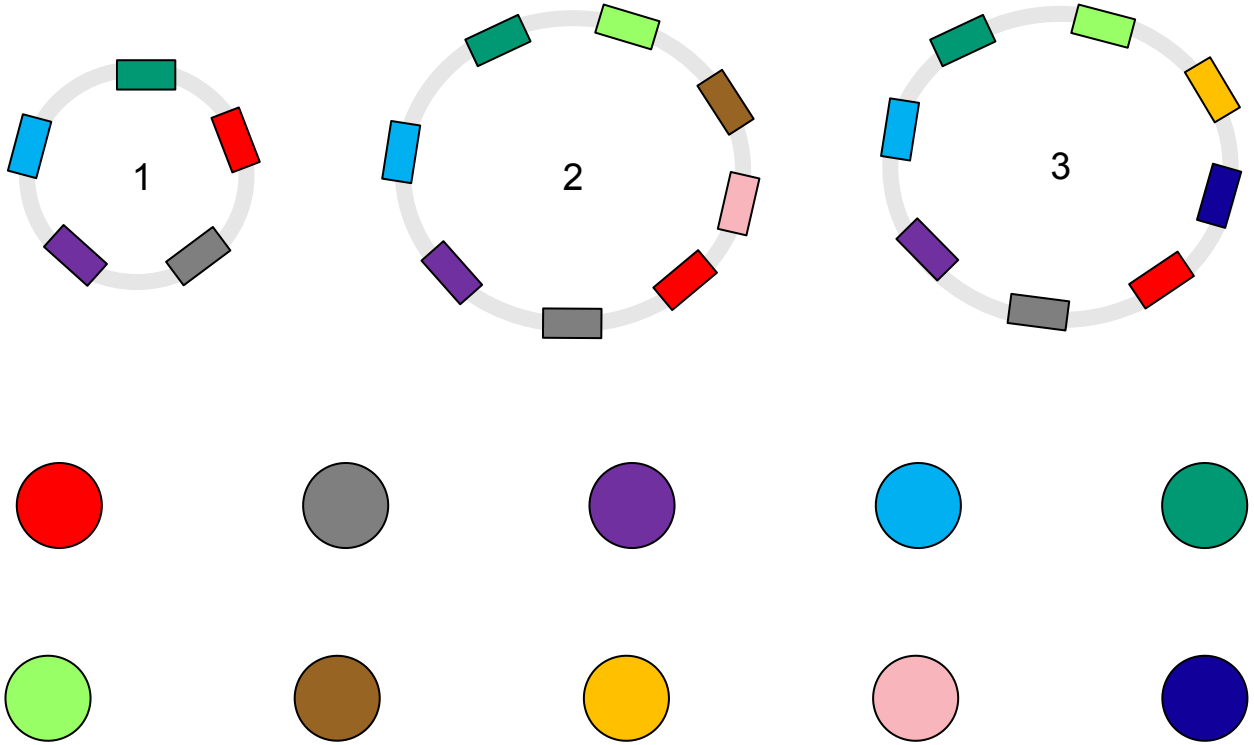
Toy example about a way to create prokaryotic pangenome



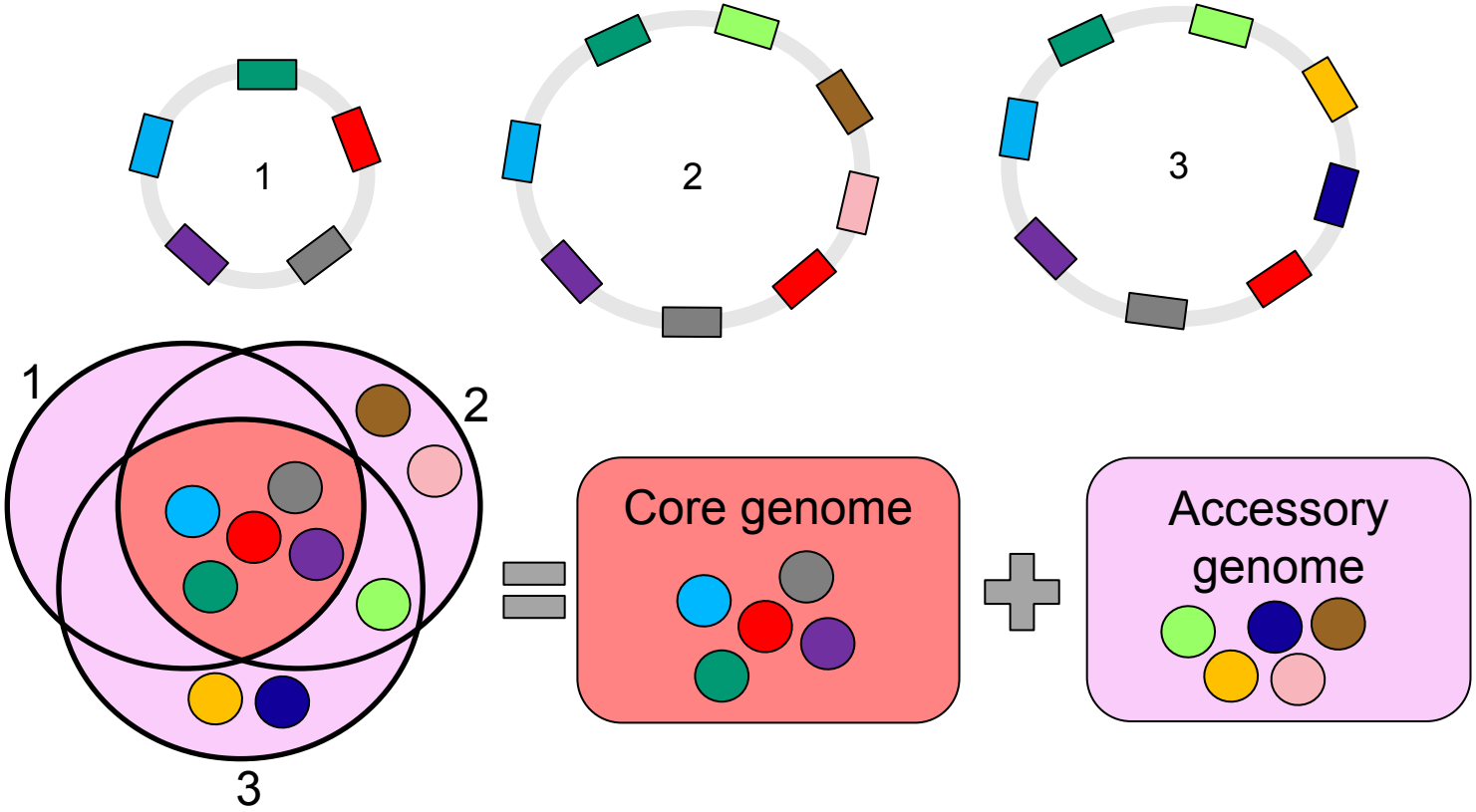
Toy example about a way to create prokaryotic pangenome



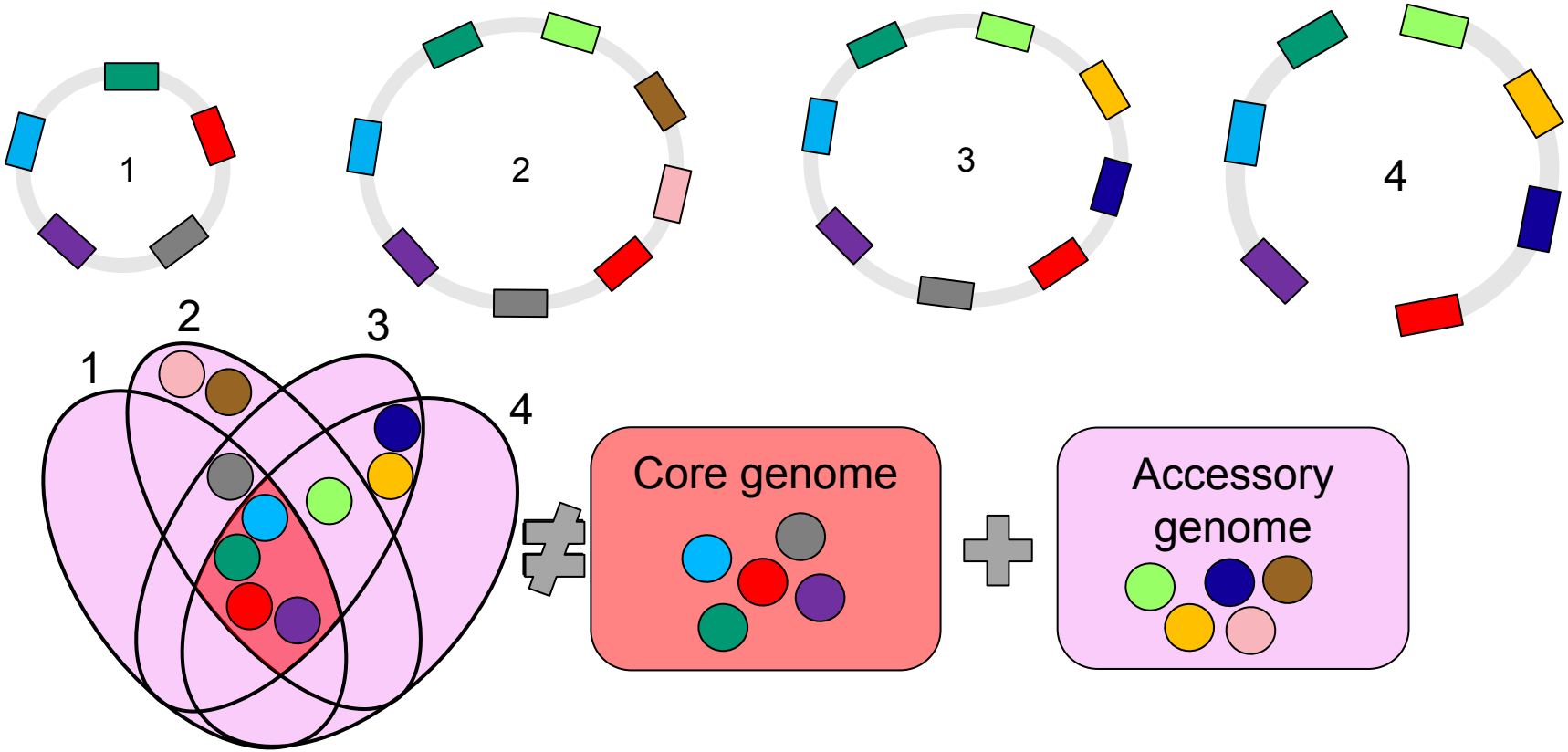
Toy example about a way to create prokaryotic pangenome



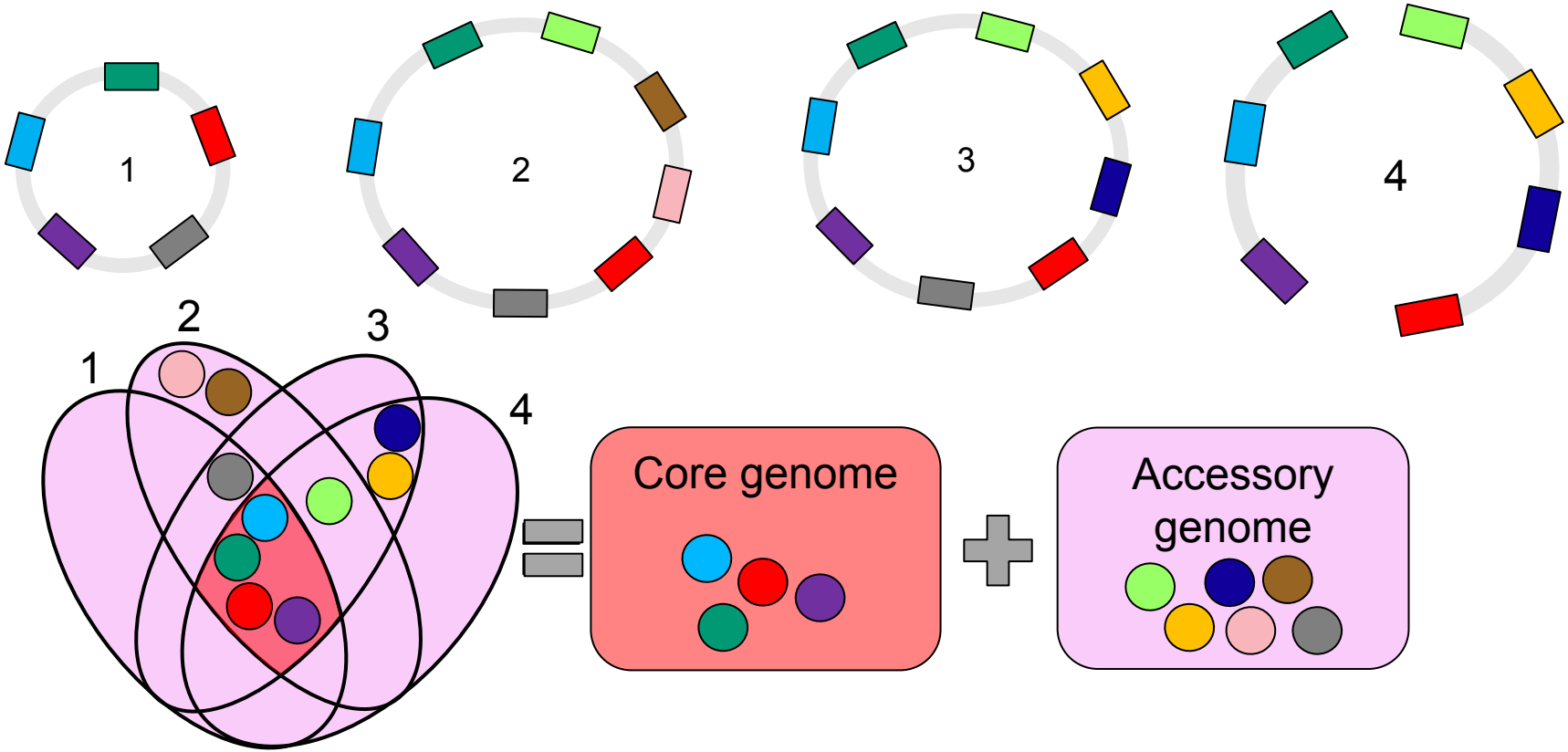
Toy example about a way to create prokaryotic pangenome



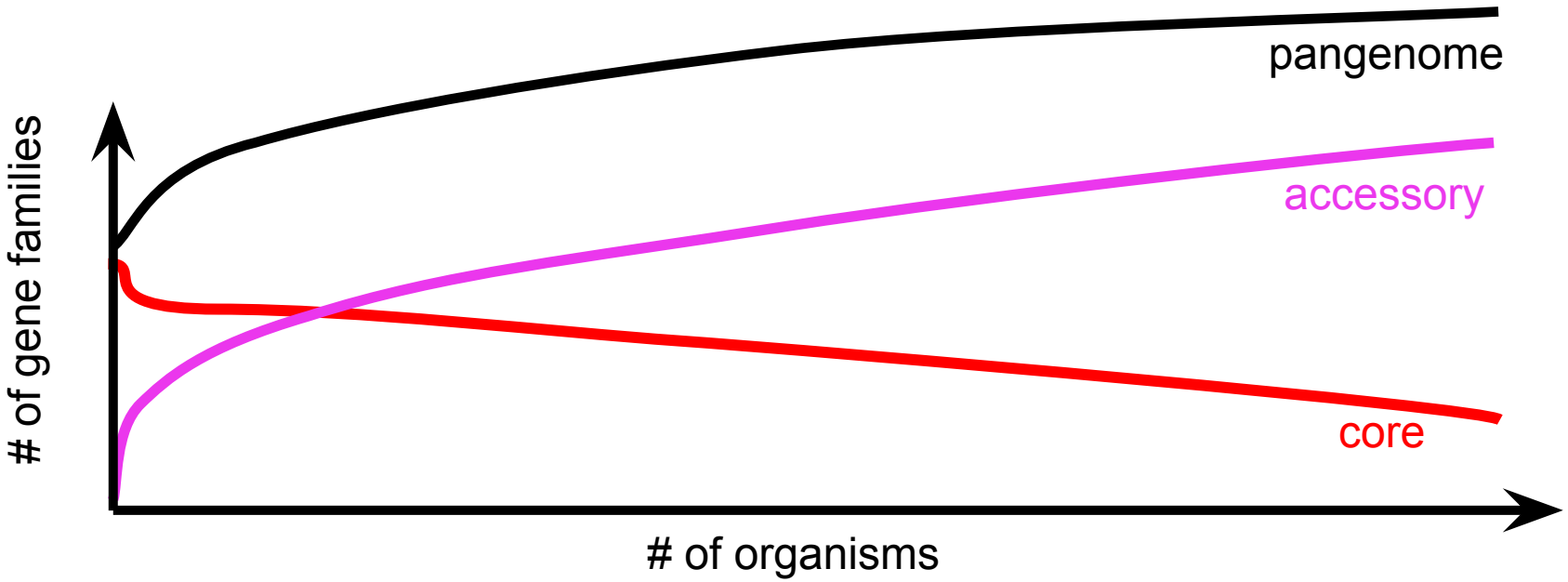
Toy example about a way to create prokaryotic pangenome



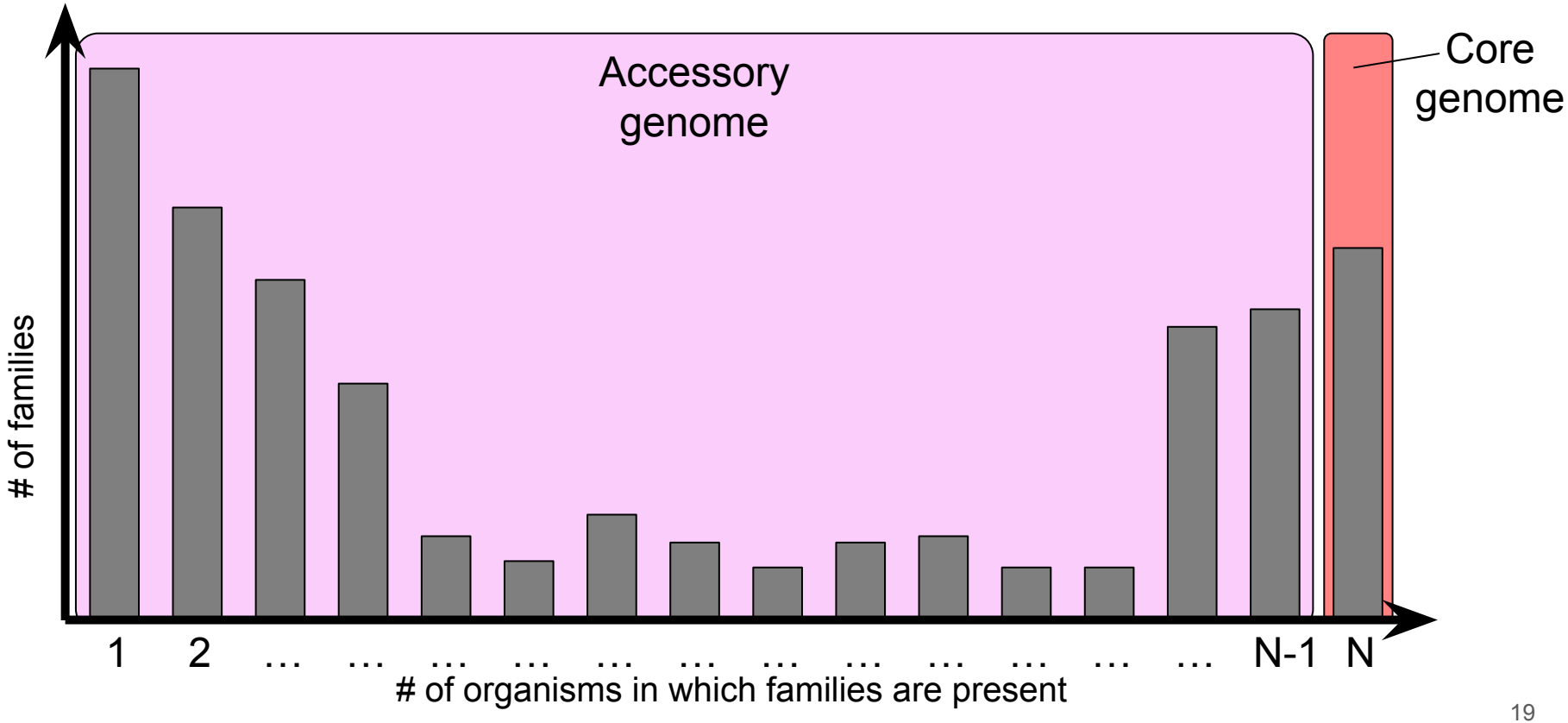
Toy example about a way to create prokaryotic pangenome



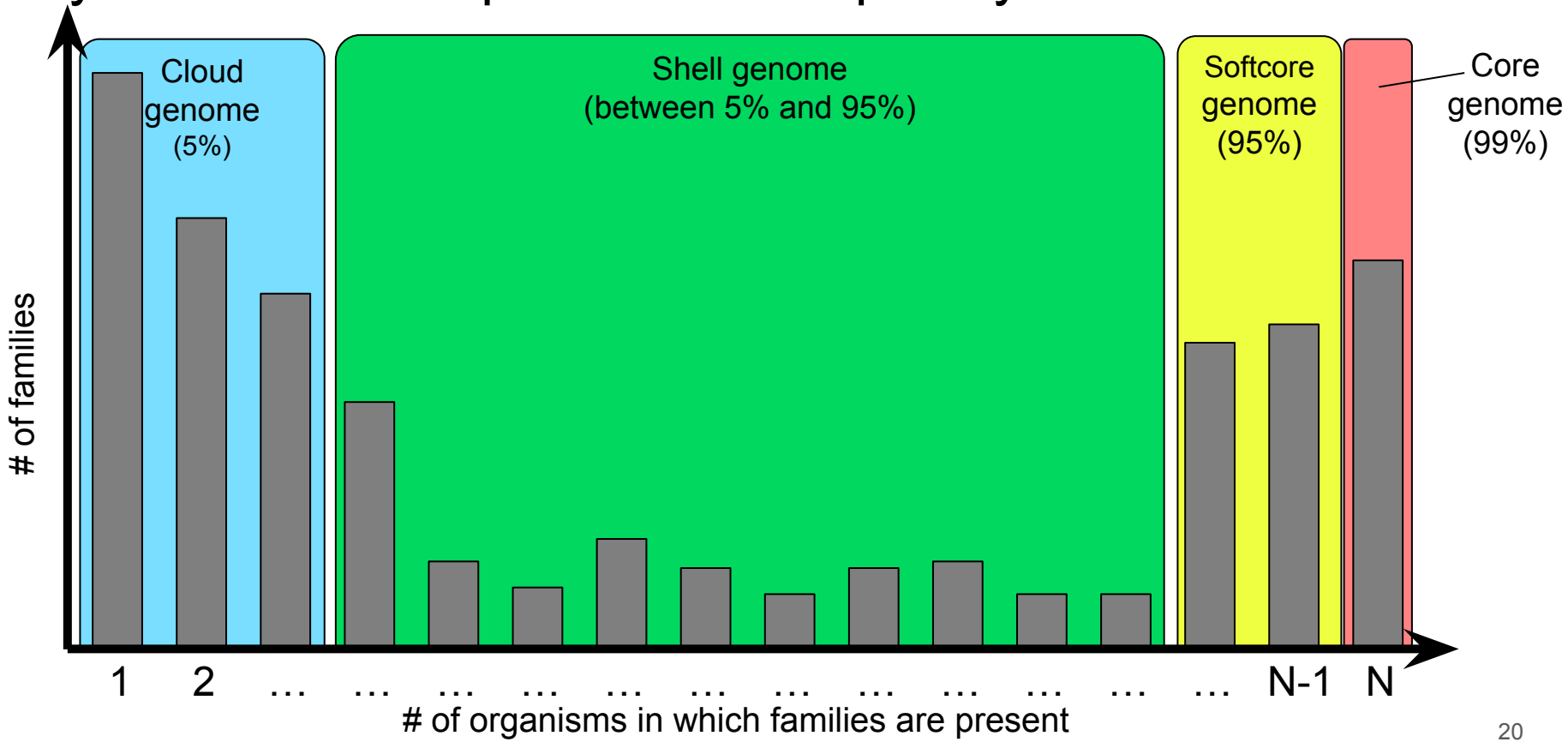
Evolution of pangenome metrics



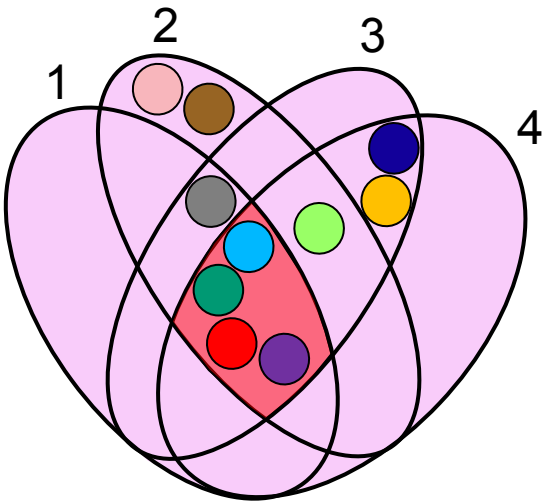
Asymmetric U-Shaped Gene Frequency Distribution



Asymmetric U-Shaped Gene Frequency Distribution



Presence/absence matrix



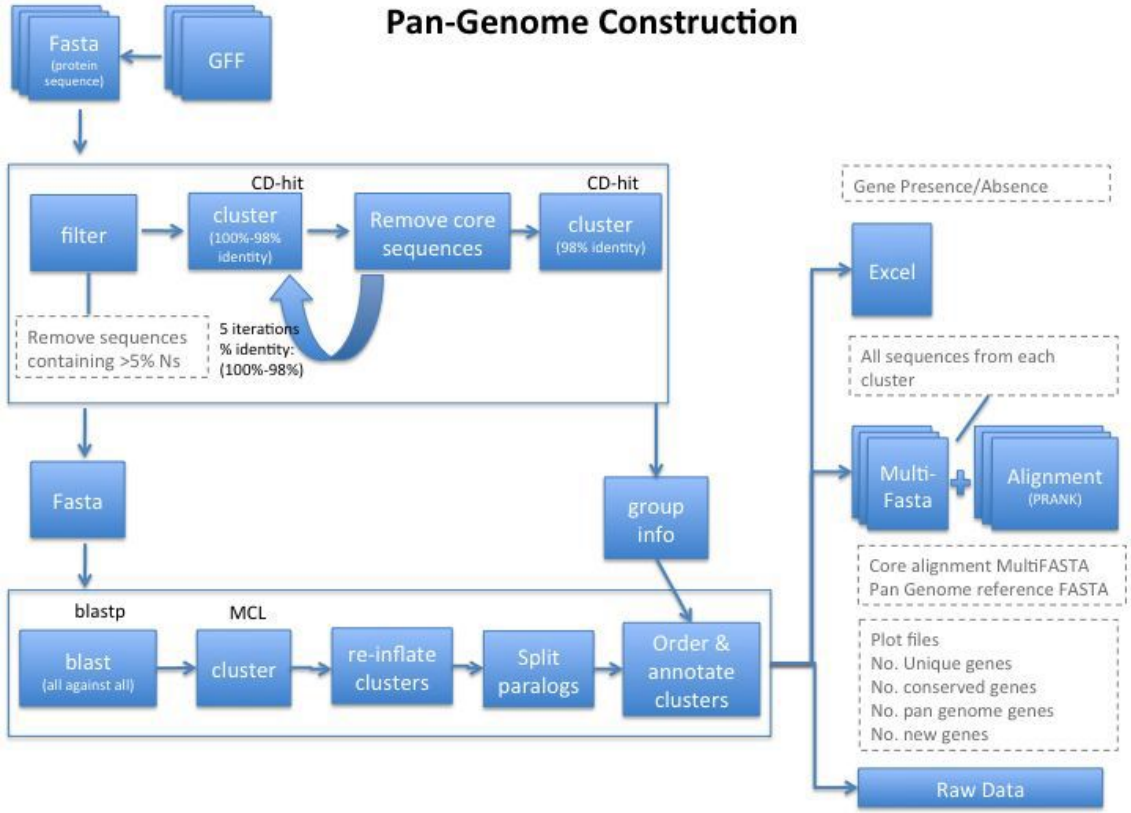
=

Presence/Absence matrix

		org1	org2	org3	org4
1		1	1	1	1
.		1	1	1	1
.		1	1	1	1
.		1	1	1	1
.		1	1	1	0
i		0	1	1	1
.		0	0	1	1
.		0	0	1	1
.		0	1	0	0
F		0	1	0	0
		1	. j .		N

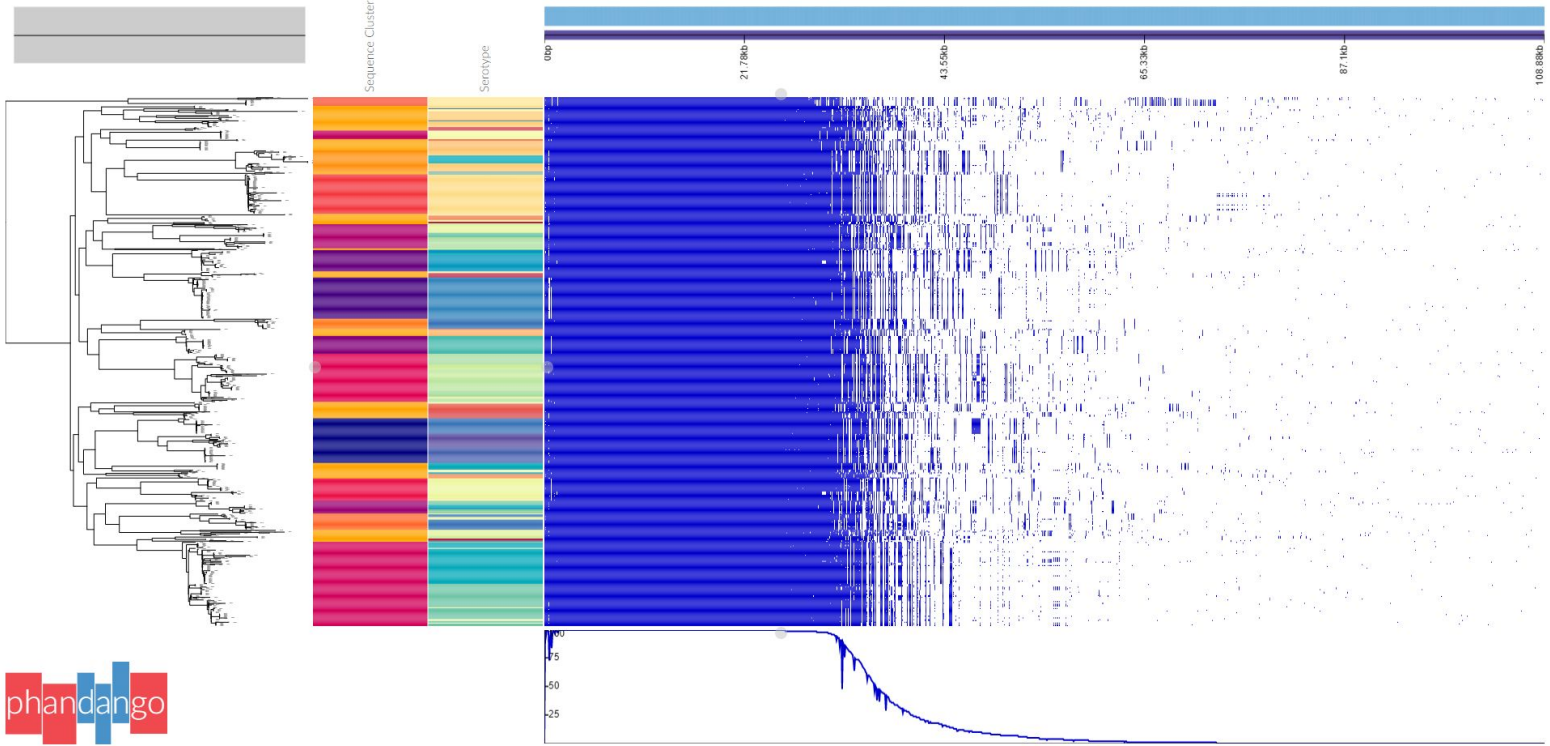
Roary to build a pangenome

- Rapid large-scale prokaryote pan genome analysis
  - Design to easily use the Prokka output
  - Based on CD-HIT to make gene families refined using BLAST



PhanDjango to easily visualize a Roary output

https://jameshadfield.github.io/phandango/#/main



Scoary (panGWAS)

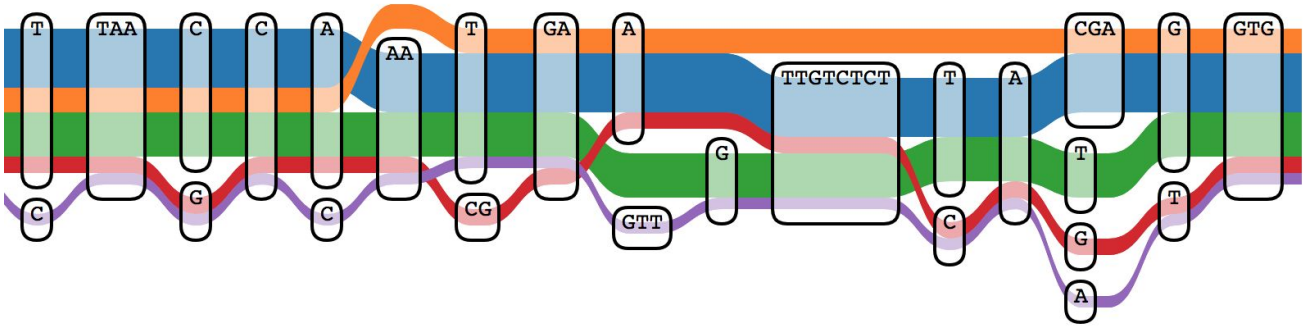
- Available on galaxy.eu (not .fr)
- Take 2 files :
  - the “gene_presence_absence.csv” file from Roary
  - A traits file (csv), example here
- It reports a list of gene families sorted by their strength of association of its presence/absence pattern to each trait

Paradigm shift

- New data model that captures the overall genomics diversity of a species

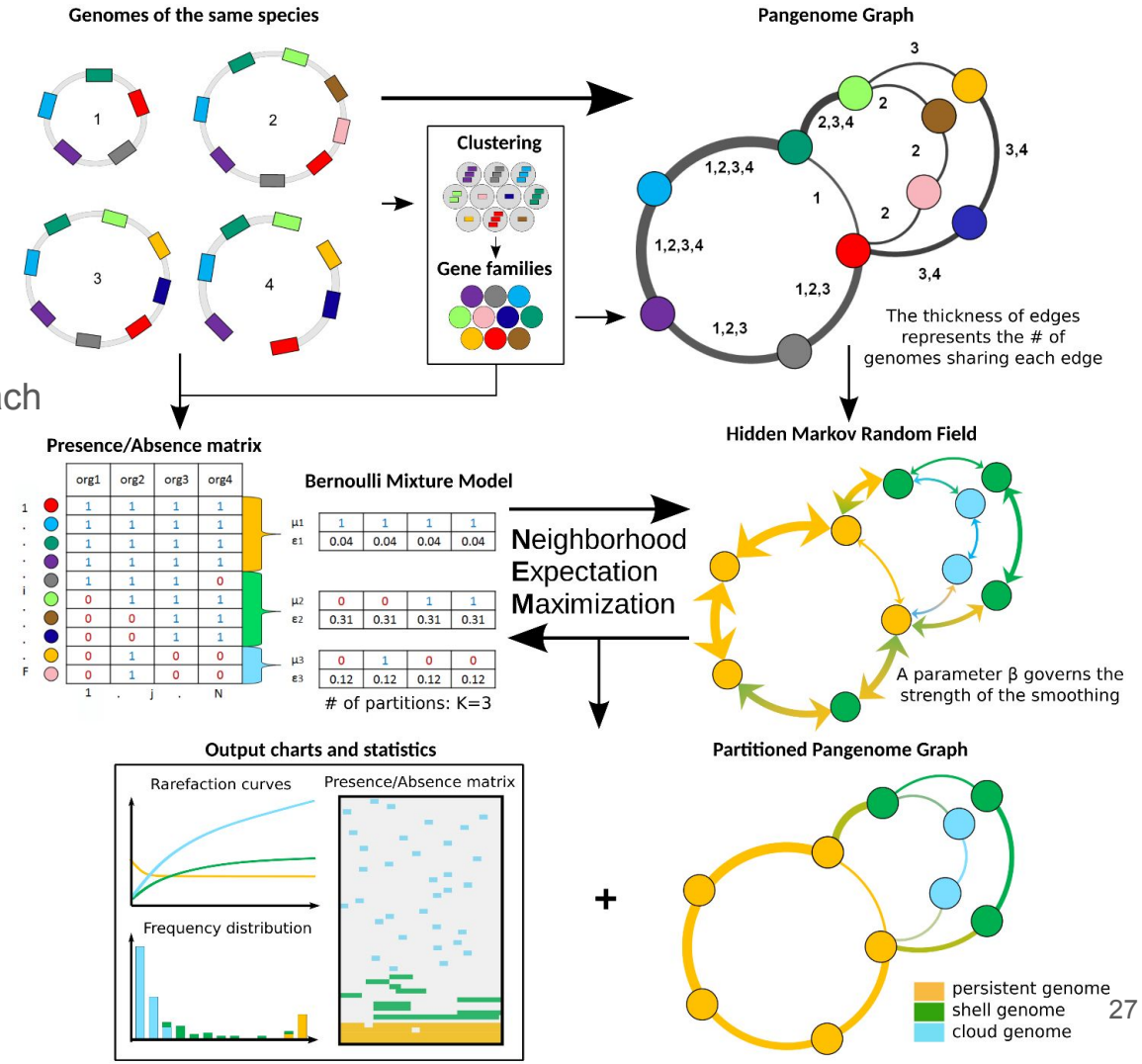
VG

- Variation Graph at sequence level
- The tools VG (https://github.com/vgteam/vg) can manage:
  - The graph data structures
  - Some (re)formating
  - Alignment,
  - Genotyping
  - Variant calling
  - Many more (very complete)…

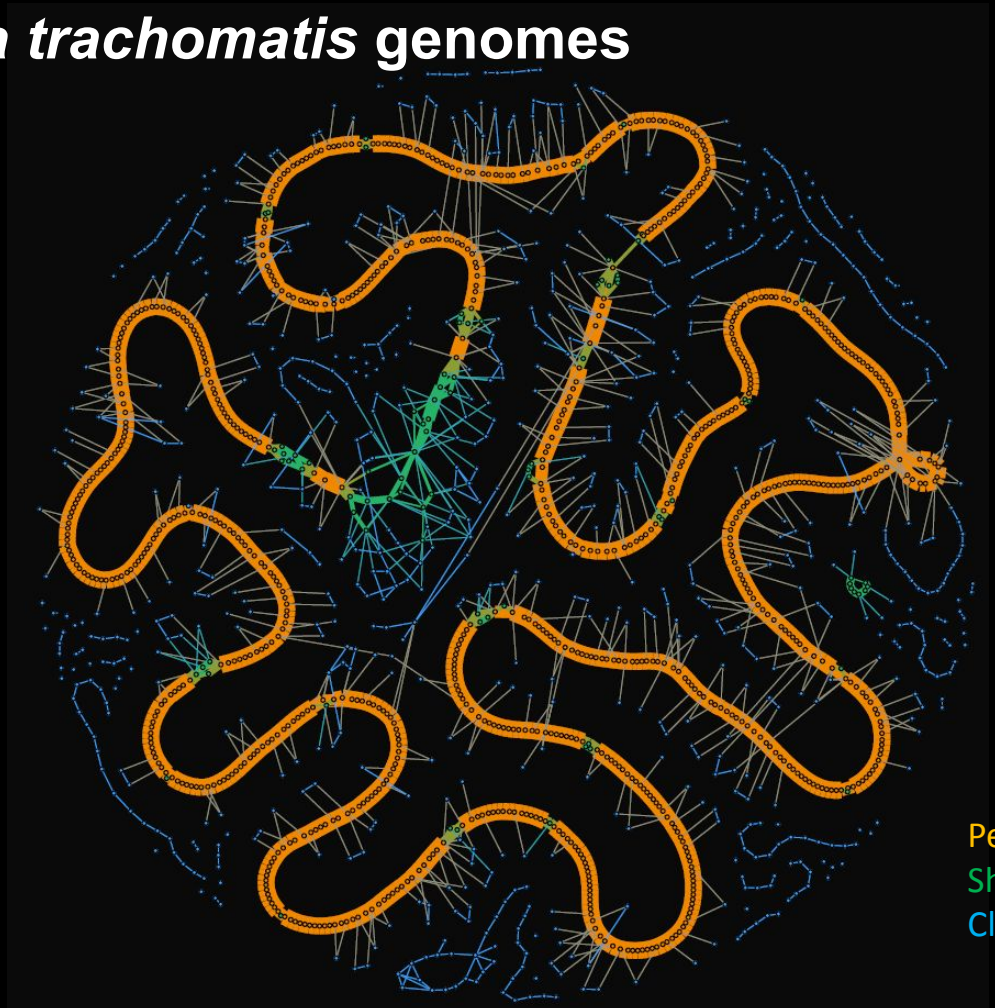


PPanGGOLiN

- A trade-off between :
  - The gene families approach
  - The sequence graph approach

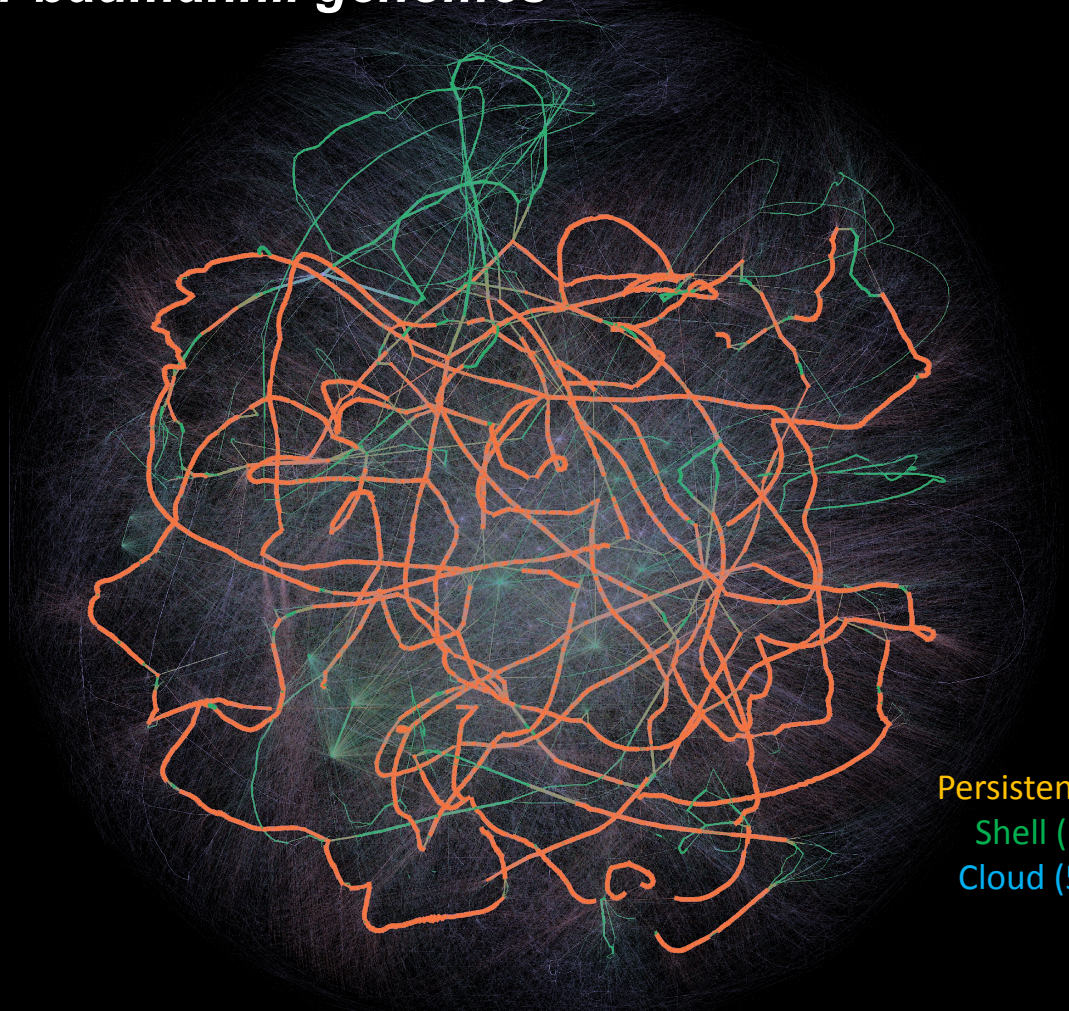


144 *Chlamydia trachomatis* genomes



Persistent (863 gene families)
Shell (84 gene families)
Cloud (761 gene families)

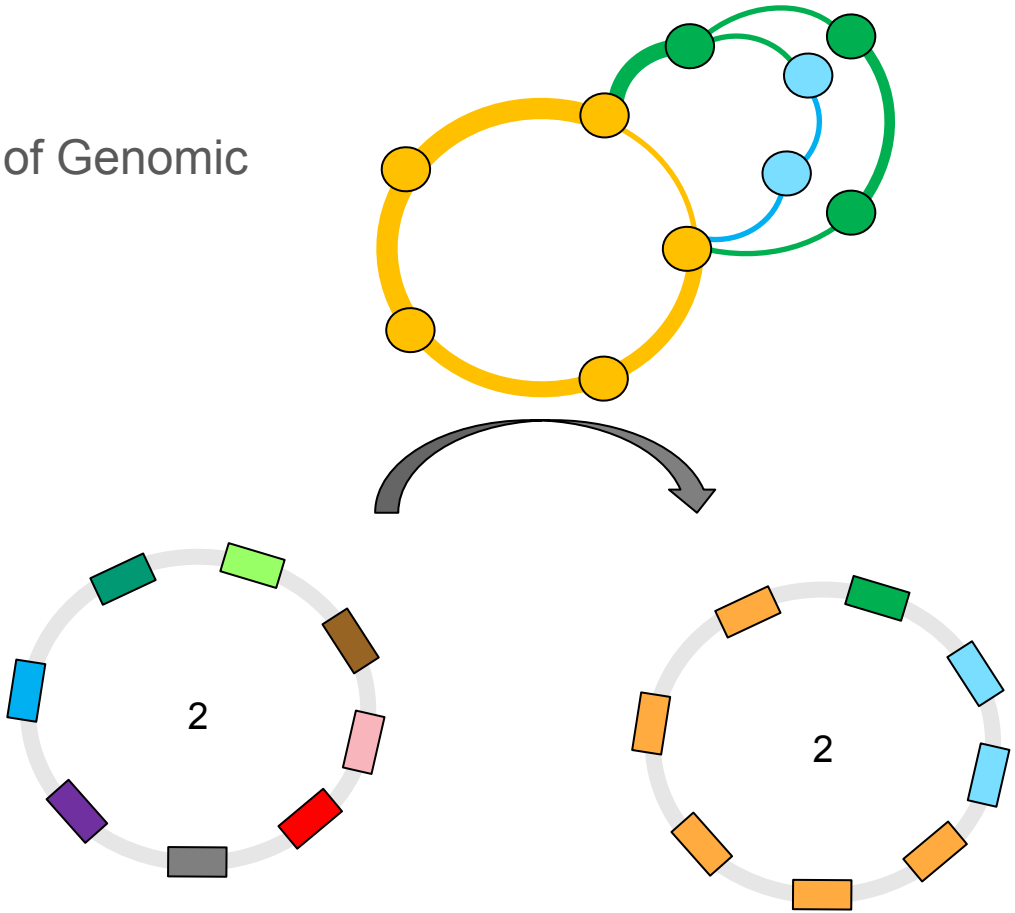
2407 *Acinetobacter baumannii genomes*



Persistent (3085 gene families)
Shell (1512 gene families)
Cloud (52812 gene families)

panRGP

Allow identifying Region of Genomic Plasticity in genomes:



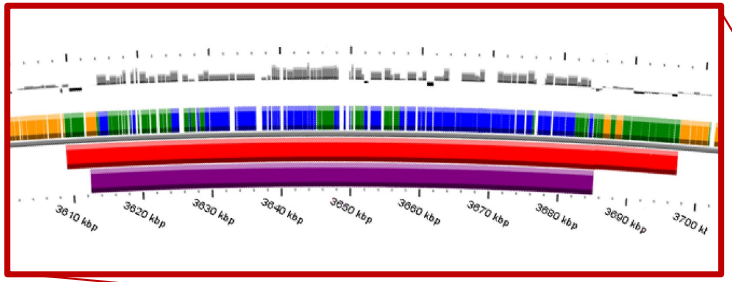
panRGP

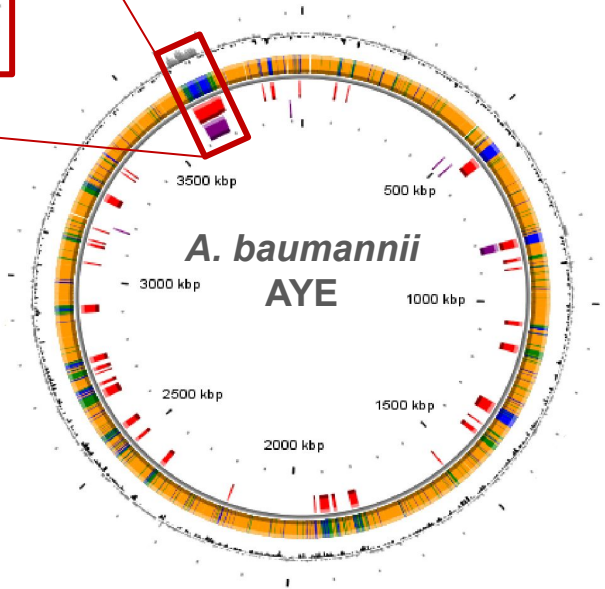
- Persistent
- Shell gene
- Cloud gene

- PanRGP prediction
- AlienHunter prediction

- %GC Variation

A. baumannii AYE - RGP35 (86kb)





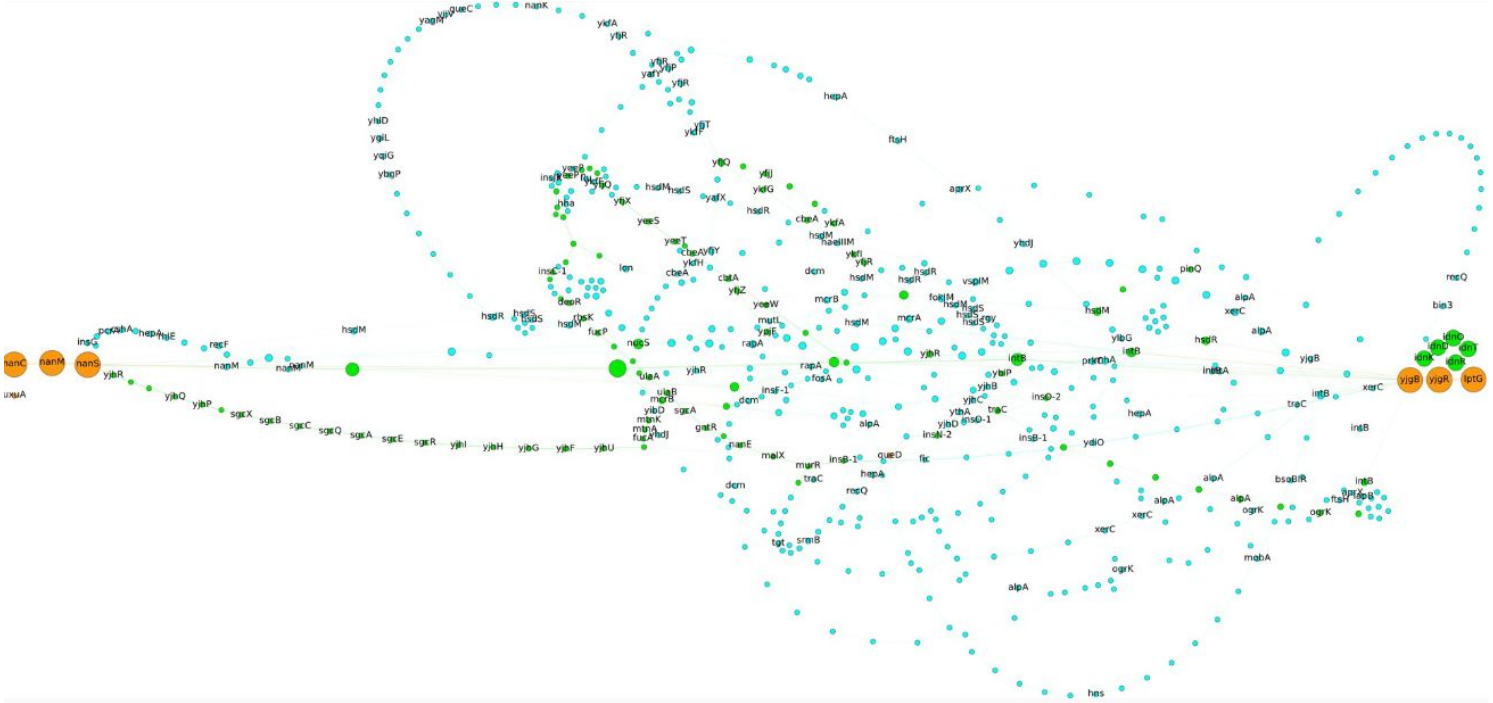
A genomic island harboring 45 genes of resistance (largest known to date)

Causing the death of 19 patients on 73 infected *(Fournier et al. 2006)*

*This feature is available in the **MicroScope** platform: https://goo.gl/ZJYVtW*

(Hot)Spot of many RGPs localized in a specific loci

Example: LeuX hotspot of RGP (*E. coli*)



Conclusions and tips

- Pangenomics is able to analysis massive datasets :
  - Useful for comparative genomics
  - Can be used as compact reference for metagenomics analysis
- Dereplication is important before doing pangenome analyses
- Use genome consistently annotated using the same pipeline to build it
- Gene diversity are not always perfectly correlated with phylogeny
- Many developments are incoming each month in the field