Pan-Genomes Are The New Reference Genomes

35th TBI Winterseminar, Feb. 2020, Bled

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Lisa-Marie Barf
lisa-marie.barf@uni-jena.de

What should a reference sequence be able to represent?

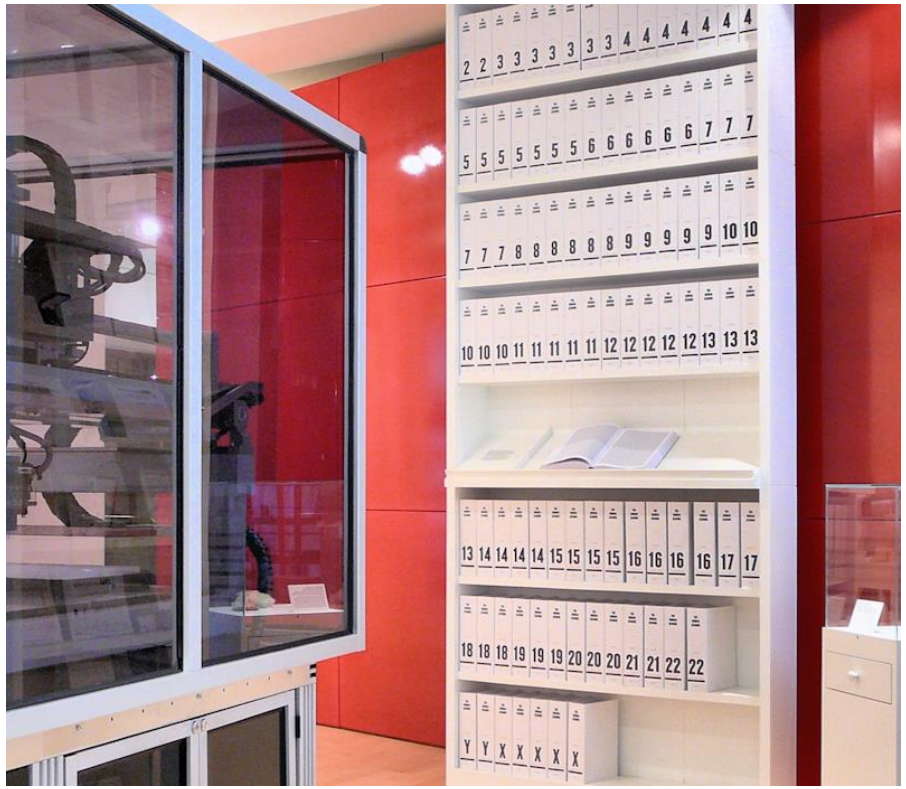
“Type strain/reference strain is usually the firstly isolated strain of the species and exhibits all of the relevant phenotypic and genotypic properties cited in the species circumscriptions.”

- Single genomes
- Functional genome
- Consensus from a population
- Maximal genome/Pan-genome

- → not one single genome as reference sequence, but rather the pan-genome

- Replace traditional linear reference genomes by richer data structures

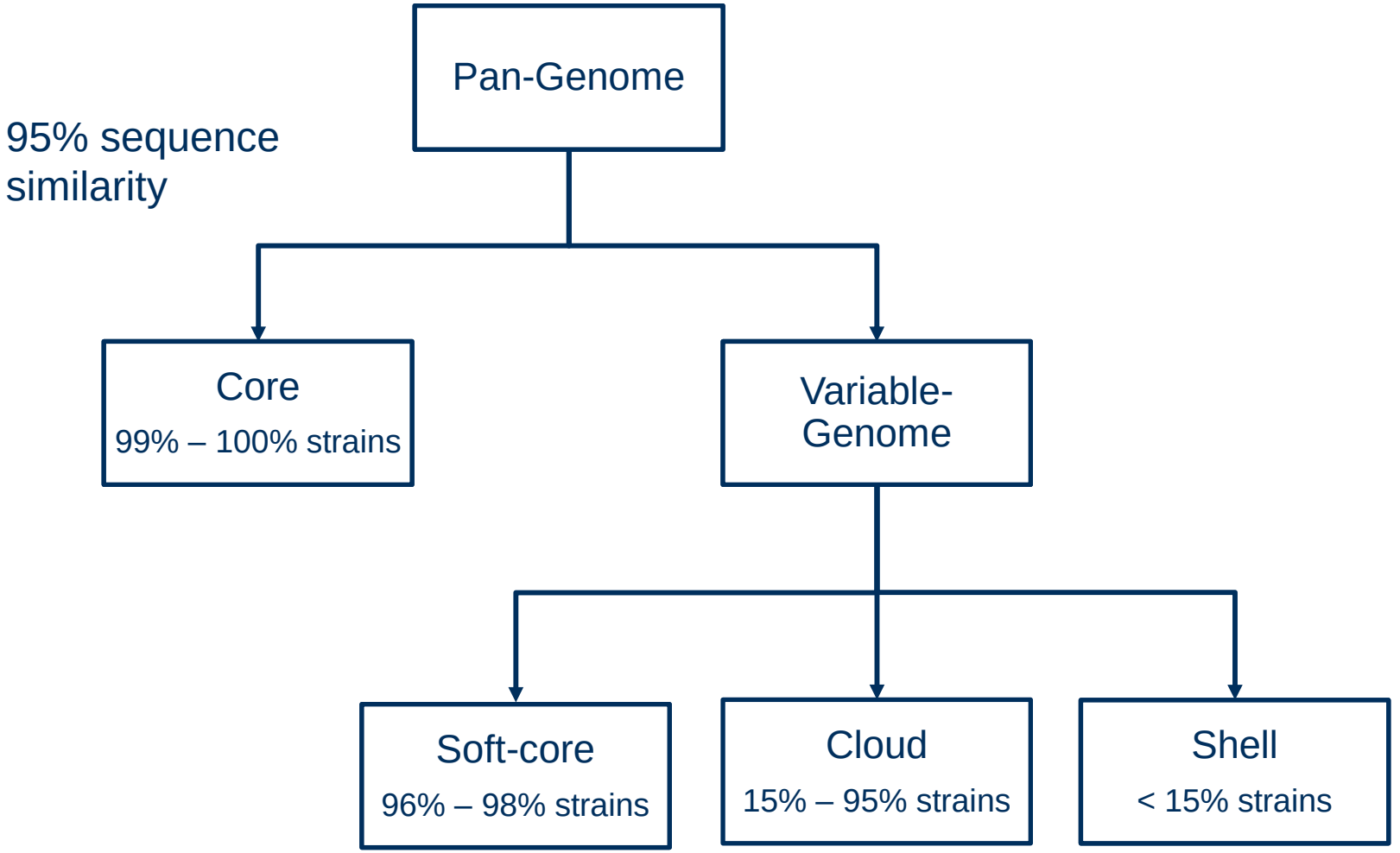
The first version of the human reference genome, 2001



https://en.wikipedia.org/wiki/Reference_genome

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

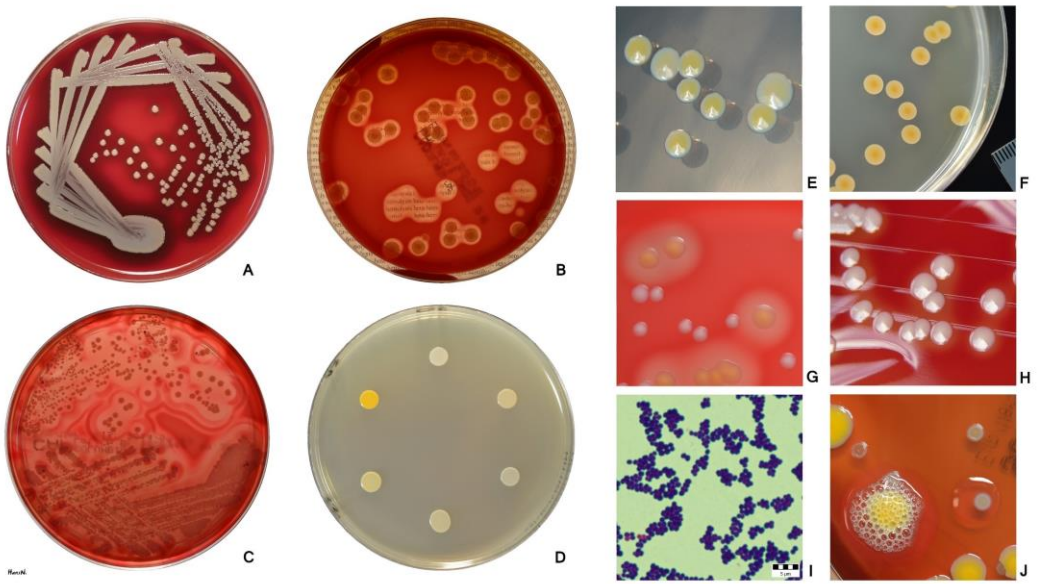
The pan-genome is a representation of all genomic content in a certain species or phylogenetic clade



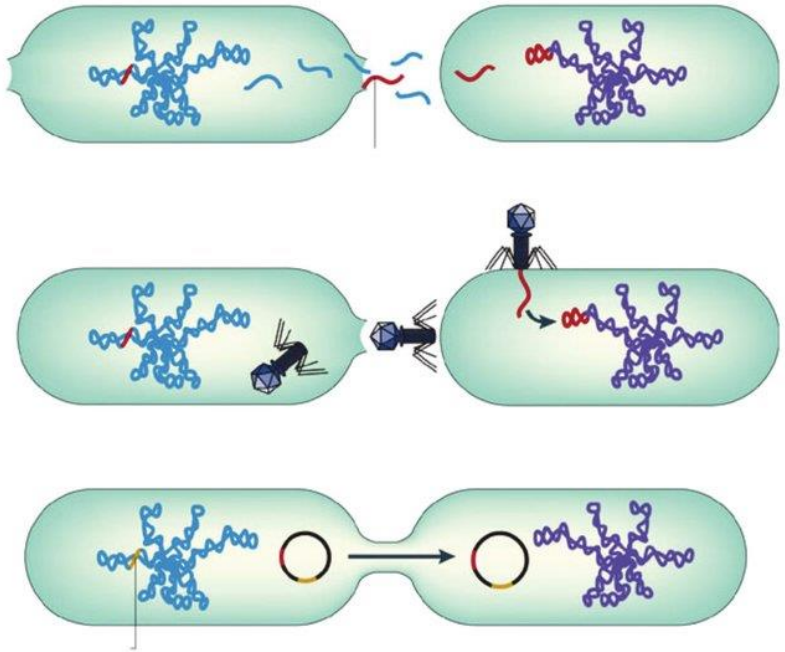
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Staphylococcus aureus: a common bacterium in the upper respiratory tract and on the skin

- Human microbiota
- Horizontal gene transfer
  → pathogenic
- Antibiotic resistance



https://en.wikipedia.org/wiki/Staphylococcus_aureus



Adapted from Furuya and Lowy, 2006

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

The power of the pan-genome: phylogeny based on the top 10 “most relevant“ *Staphylococcus aureus* strains

1	>AP017922.1 Staphylococcus aureus DNA, strain: JP080
2	>CP013231.1 Staphylococcus aureus strain UTSW MRSA 55
3	>CP038021.1 Staphylococcus aureus strain 04-002
4	>CP038268.1 Staphylococcus aureus strain O55 isolate B118
5	>CP038819.1 Staphylococcus aureus strain O82
6	>CP039848.1 Staphylococcus aureus strain 2030RH1
7	>CP040623.1 Staphylococcus aureus strain D592-HR
8	>CP040801.1 Staphylococcus aureus strain S15
9	>LN626917.1 Staphylococcus aureus strain ILRI_Eymole1/1
10	>NC_002951.2 Staphylococcus aureus strain COL

Genotyping approaches:

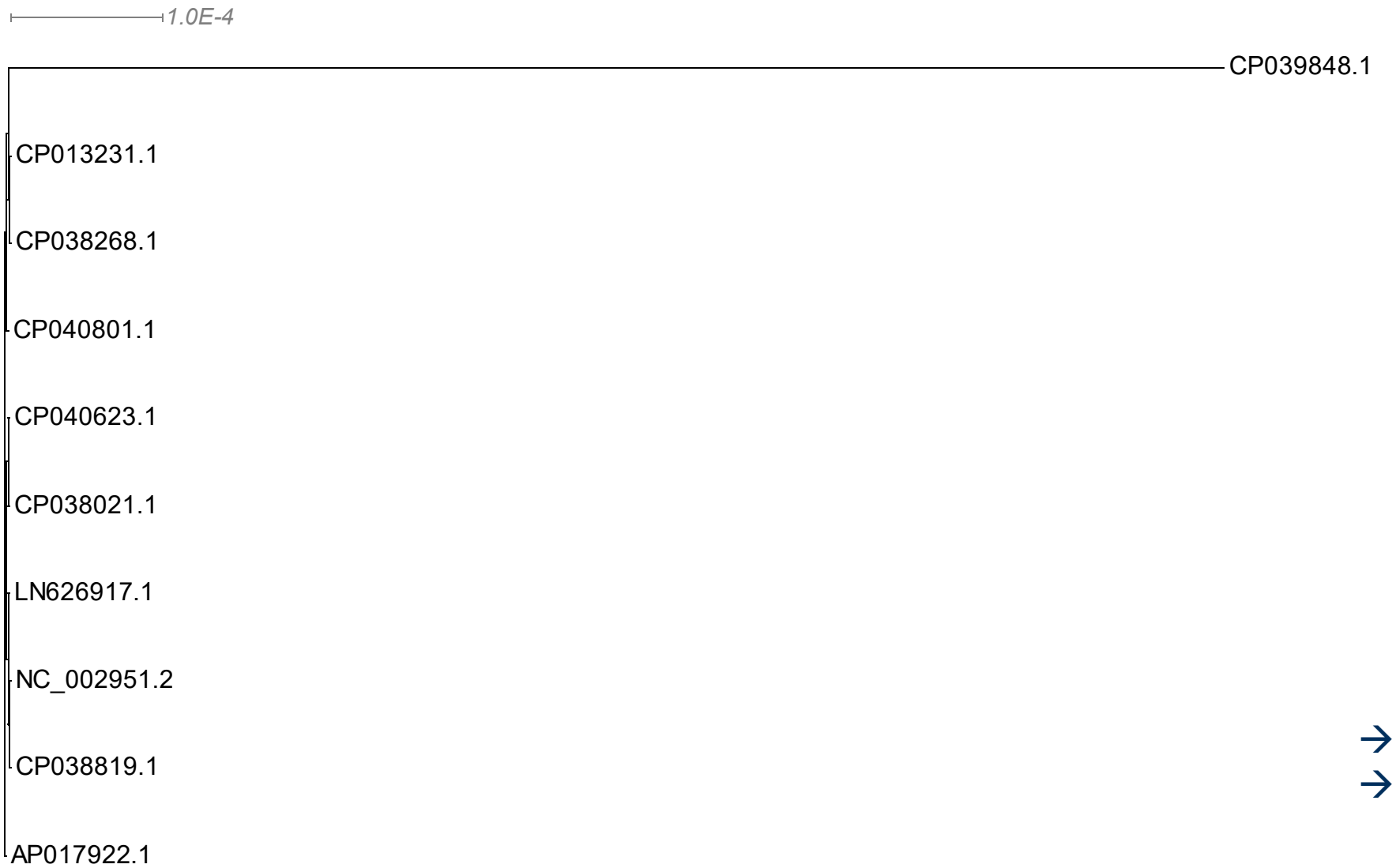
- 16S rRNA
→ one gene

- MLST (Multilocus sequence typing)
→ 7 housekeeping genes

- Core-genome
→ 1991 core genes

***Staphylococcus aureus* phylogeny**

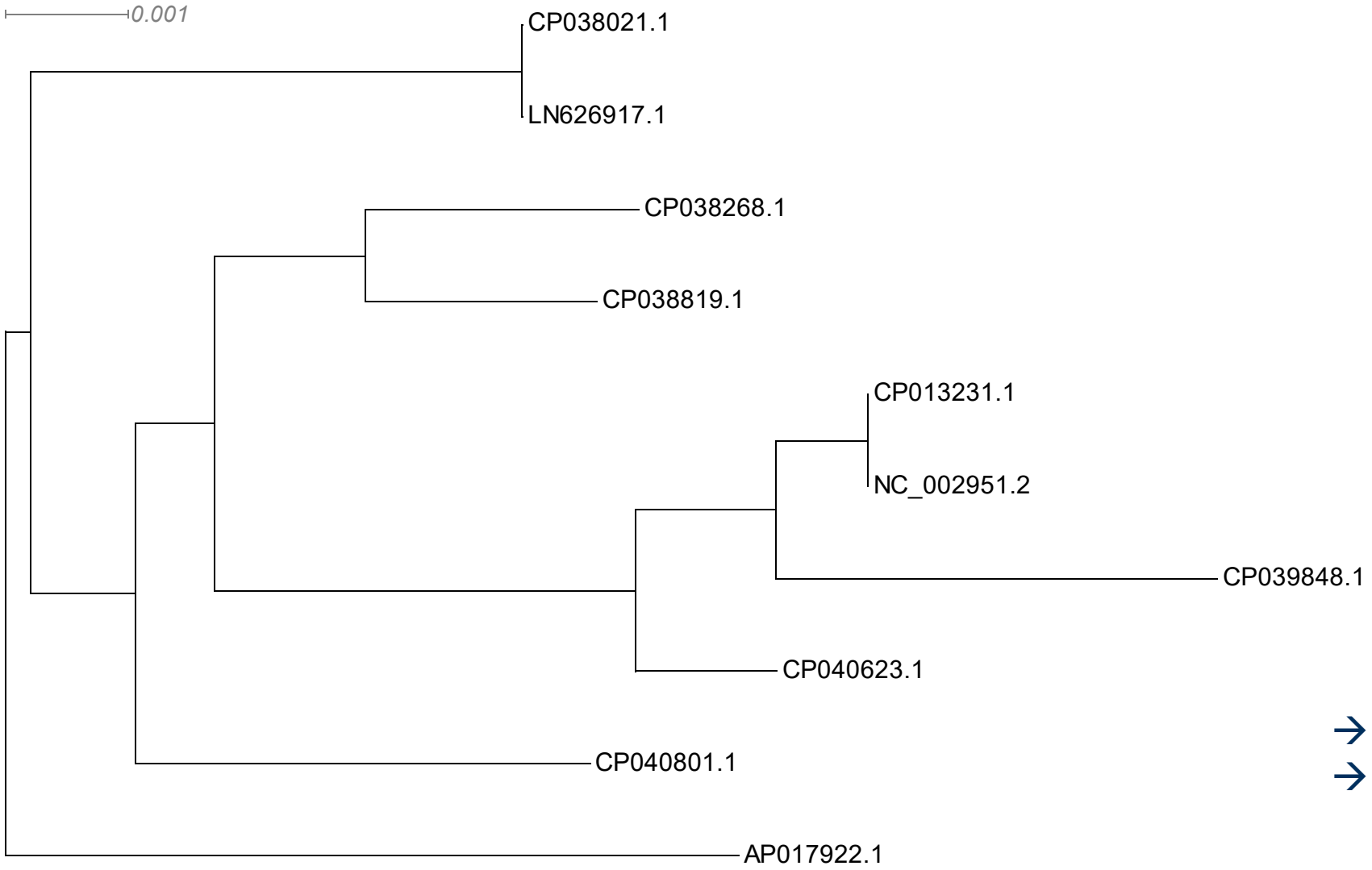
Based on: 16S rRNA (1 gene)



→ Closely related strains
→ One outlier

Staphylococcus aureus **phylogeny**

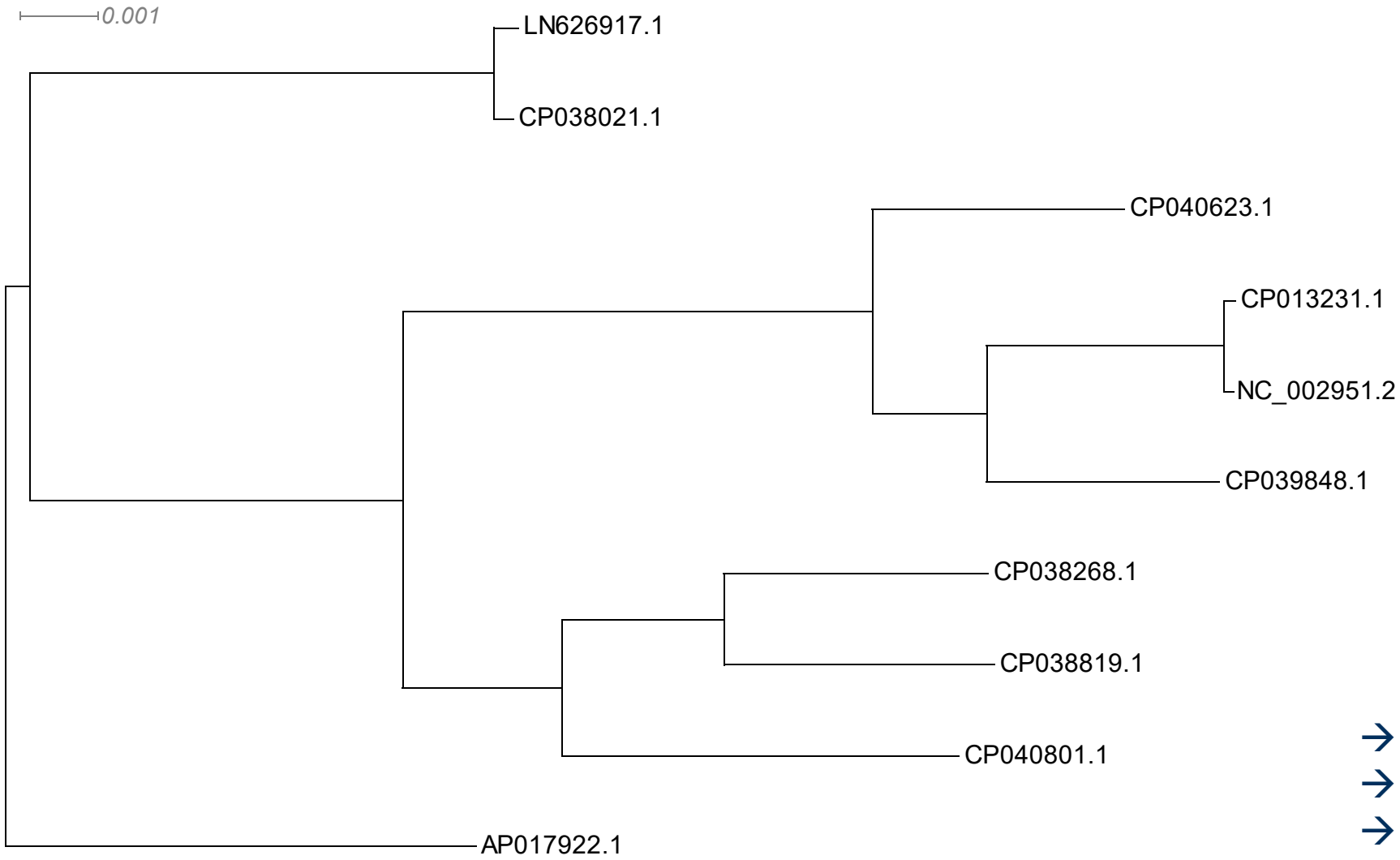
Based on: MLST (7 housekeeping genes)



→ Better resolution
→ Still identical strains

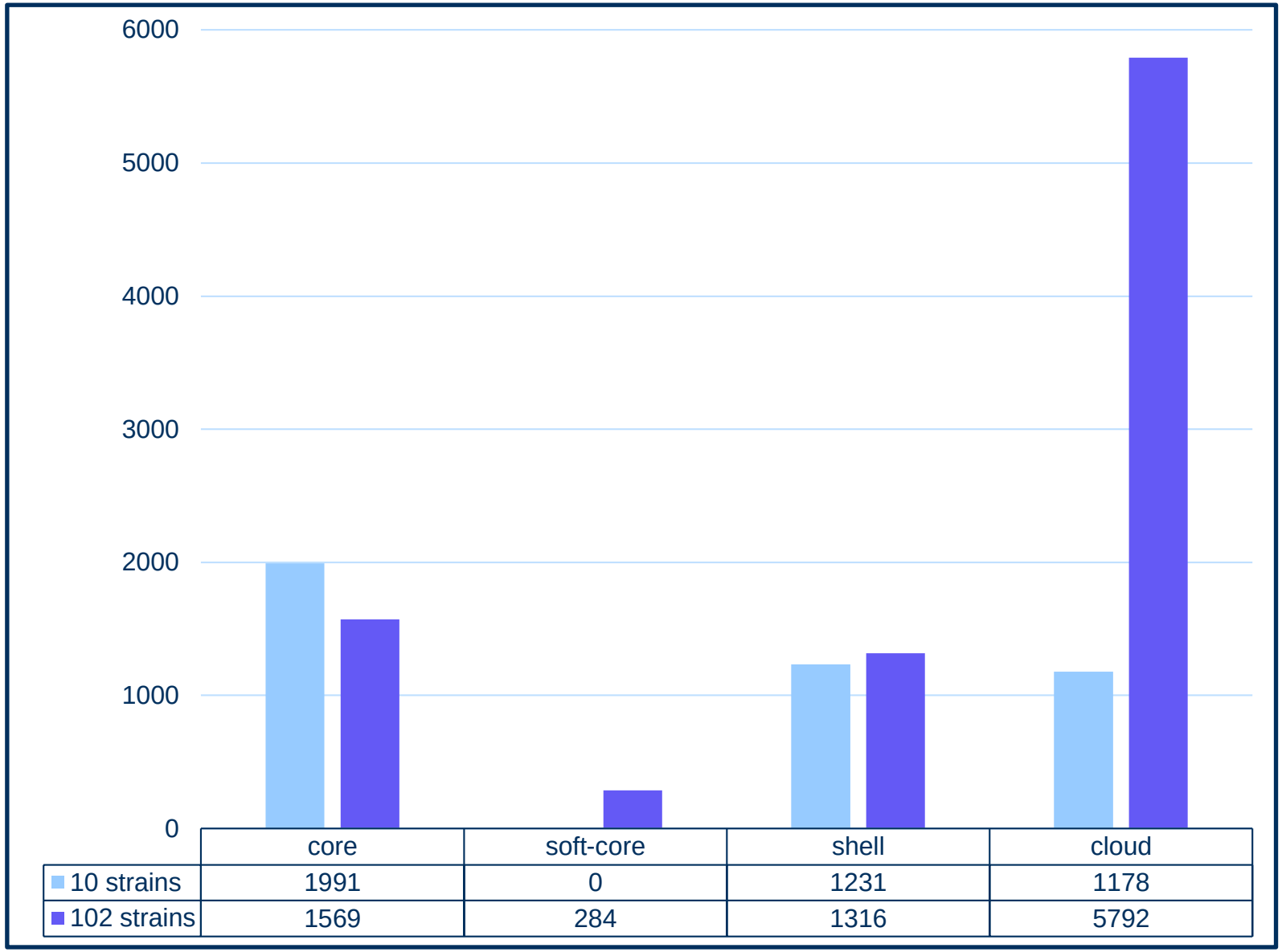
Staphylococcus aureus phylogeny

Based on: Core-genome (1991 genes)



→ No identical
→ Changed topology
→ Robust?

The *Staphylococcus aureus* core-genome is robust: 10 vs 102 strains



	core	soft-core	shell	cloud
10 strains	1991	0	1231	1178
102 strains	1569	284	1316	5792

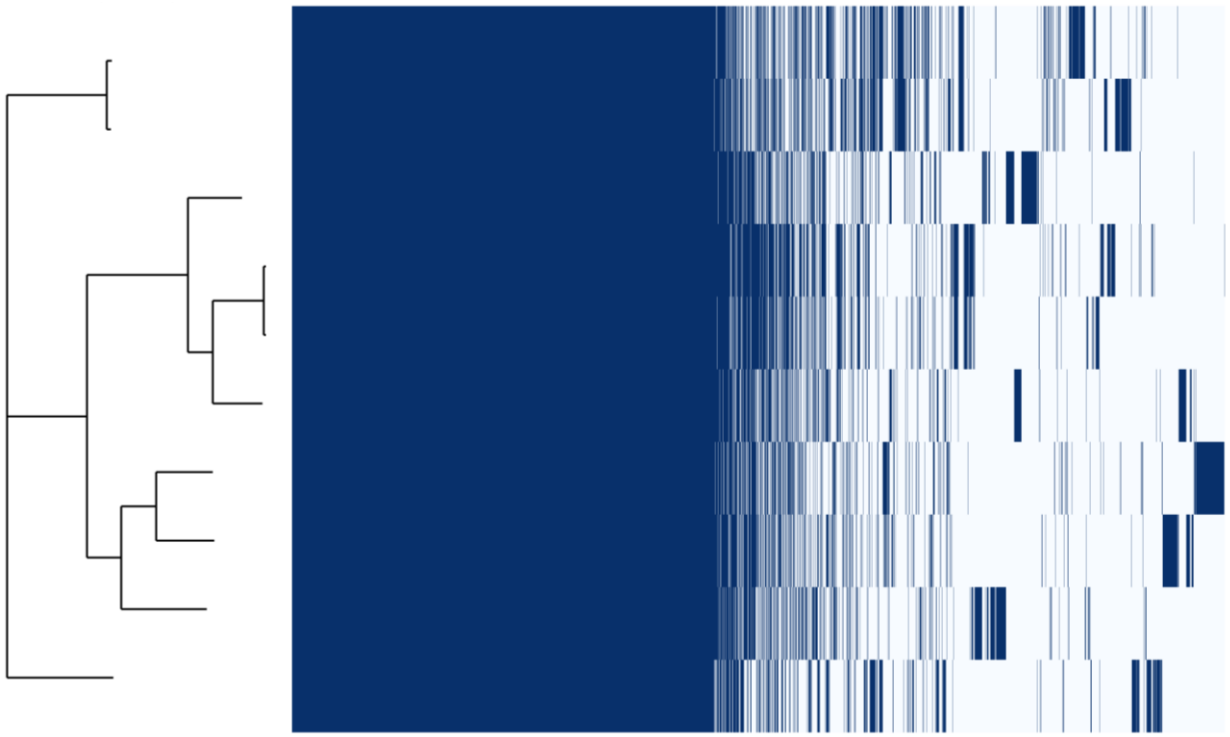
→ Robust Core-genome

Core: 99% – 100%; soft-core: 96% – 98%; shell: 15% – 95%; cloud: < 15%

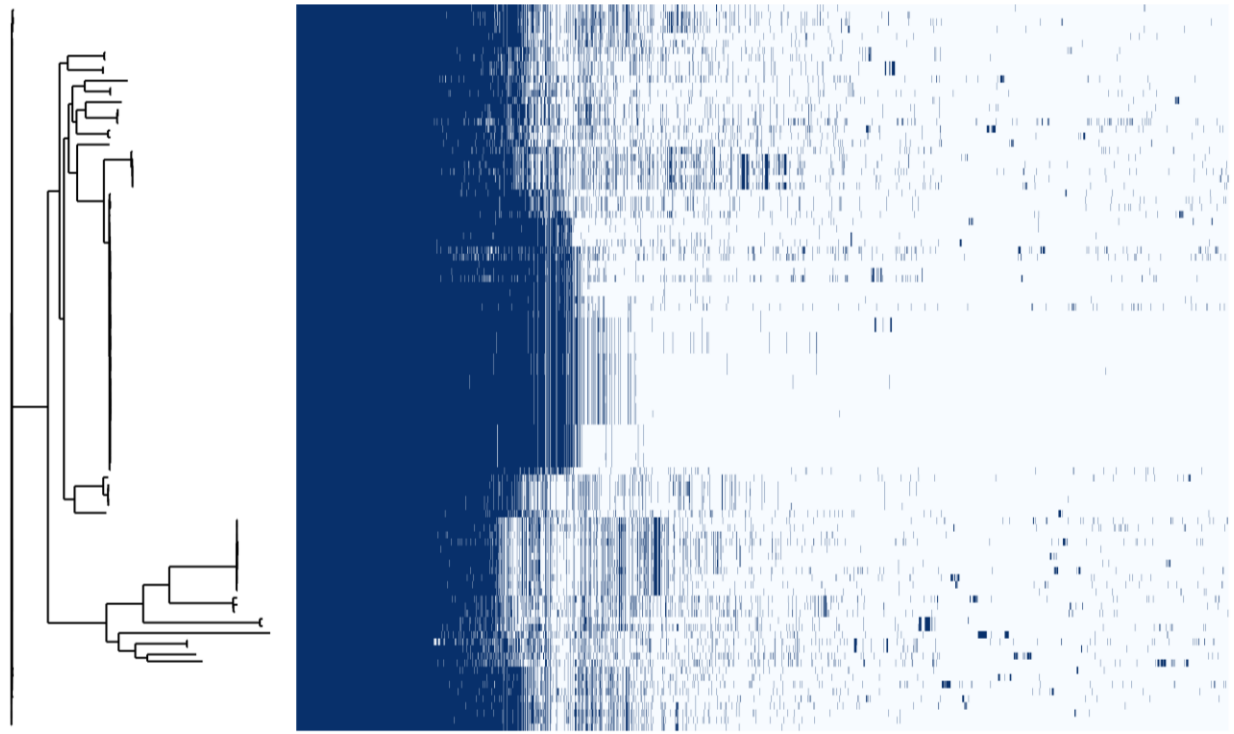
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

The *Staphylococcus aureus* core-genome is robust: 10 vs 102 strains

Pan-genome matrix
10 strains



Pan-genome matrix
102 strains

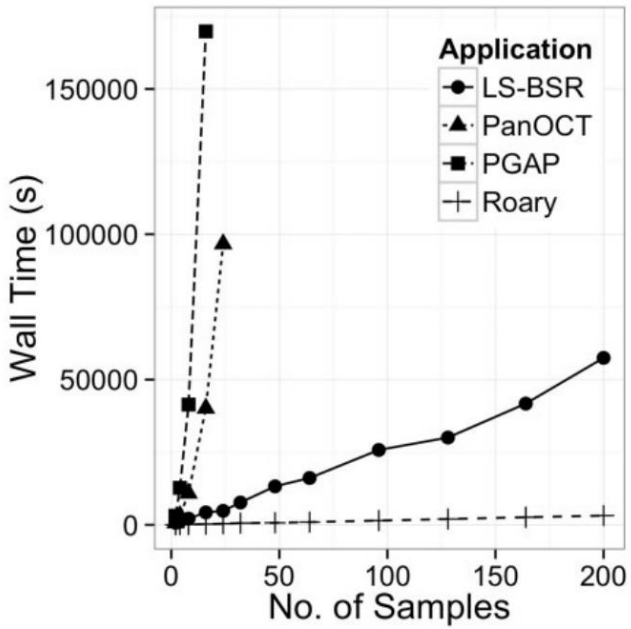


FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Next step: The *Staphylococcus aureus* pan-genome

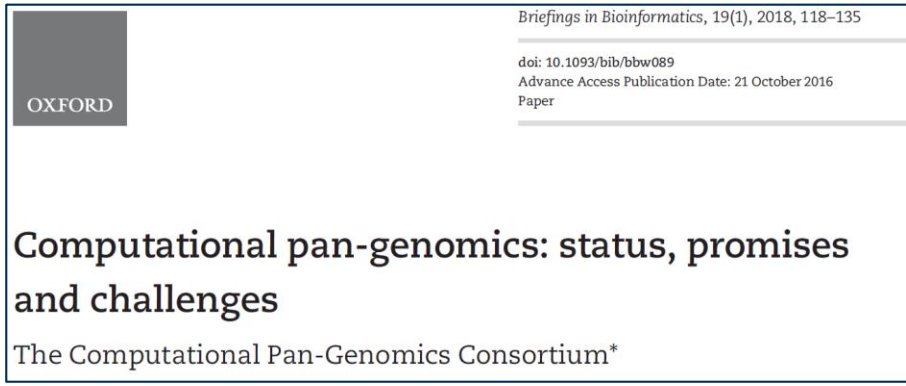
- 40,000 strains
- A single reference strain is not representative for a whole species
- Use pan-genome to characterize species

- Runtime
- Storage
- Computational solution



FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Outlook - The *Staphylococcus aureus* pan-genome
Computational challenges



Data structures
Design goals:

- Construction and maintenance
- Coordinate system
- Biological features and computational layers
- Data retrieval
- Searching
- Comparing

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Outlook - The *Staphylococcus aureus* pan-genome
Computational challenges

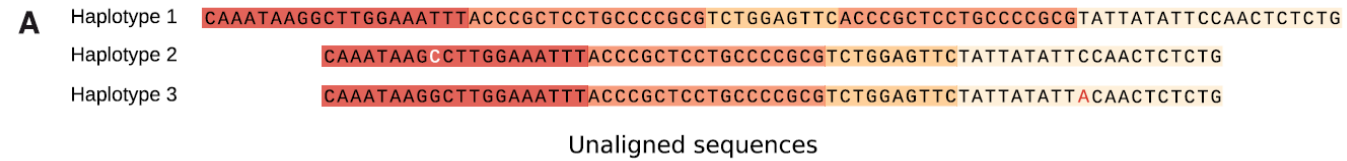
Data structure
Approaches

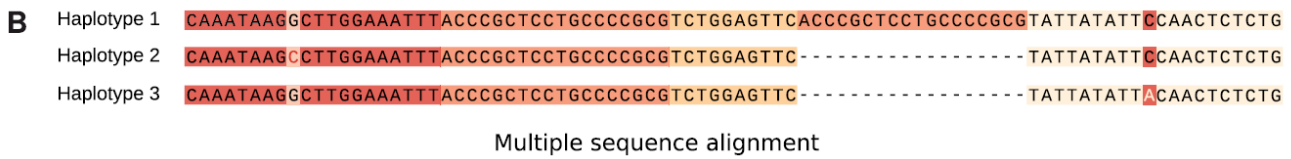
Briefings in Bioinformatics, 19(1), 2018, 118–135

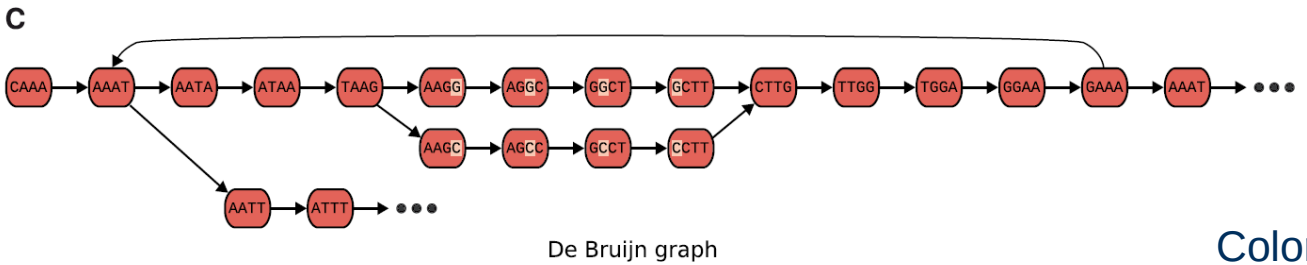
doi: 10.1093/bib/bbw089
Advance Access Publication Date: 21 October 2016
Paper

Computational pan-genomics: status, promises and challenges

The Computational Pan-Genomics Consortium*







Colored De Bruijn

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Outlook - The *Staphylococcus aureus* pan-genome
Computational challenges

Variant calling and genotyping approaches

- Difference between newly sequenced genome and a reference

Visualization

- View large genome sets
- Homology relationships

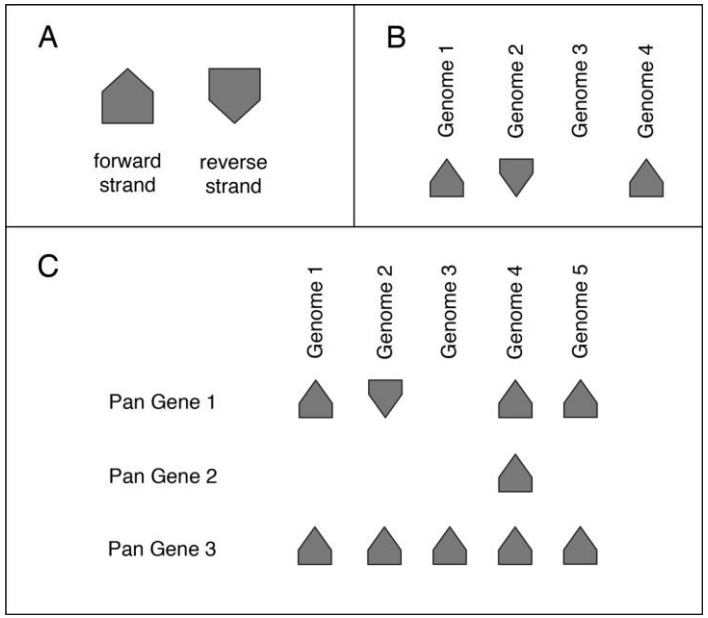
OXFORD

Briefings in Bioinformatics, 19(1), 2018, 118–135

doi: 10.1093/bib/bbw089
Advance Access Publication Date: 21 October 2016
Paper

Computational pan-genomics: status, promises and challenges

The Computational Pan-Genomics Consortium*



Pan-Tetris, Henning et al., 2015

Conclusion

- A single reference genome is not feasible to represent a whole species, but rather the pan-genome.
- Move away from linear reference genomes towards reference systems (graph based).
- Solve computational challenges in terms of storage and visualization.

Many thanks to:

RNA BIOINFORMATICS & HIGH-THROUGHPUT ANALYSIS

Leibniz | ipht

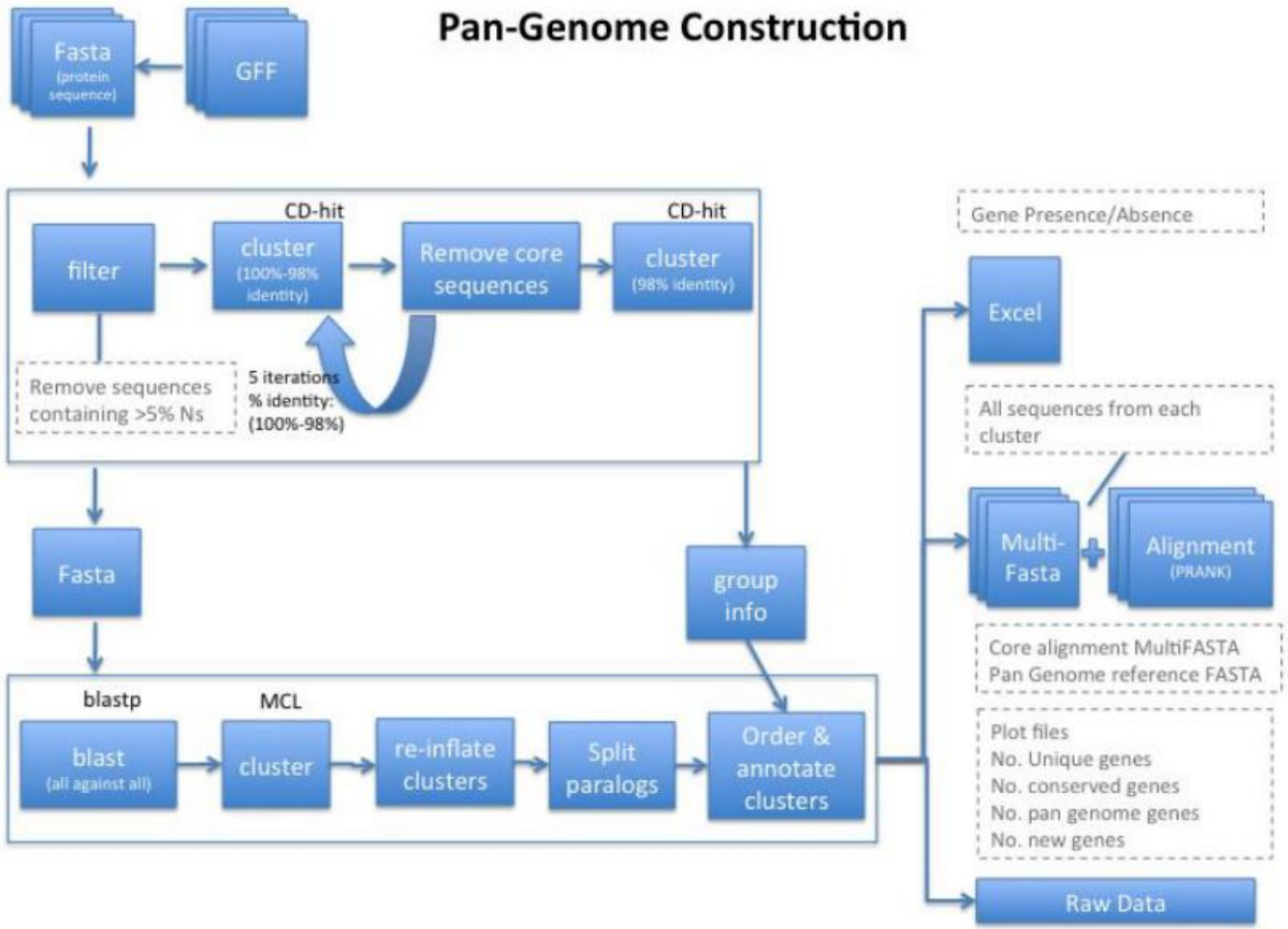
Max-Planck-Institut
für Menschheitsgeschichte
Max Planck Institute
for the Science of Human History

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Thank you for your attention!

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

2.2 Method description

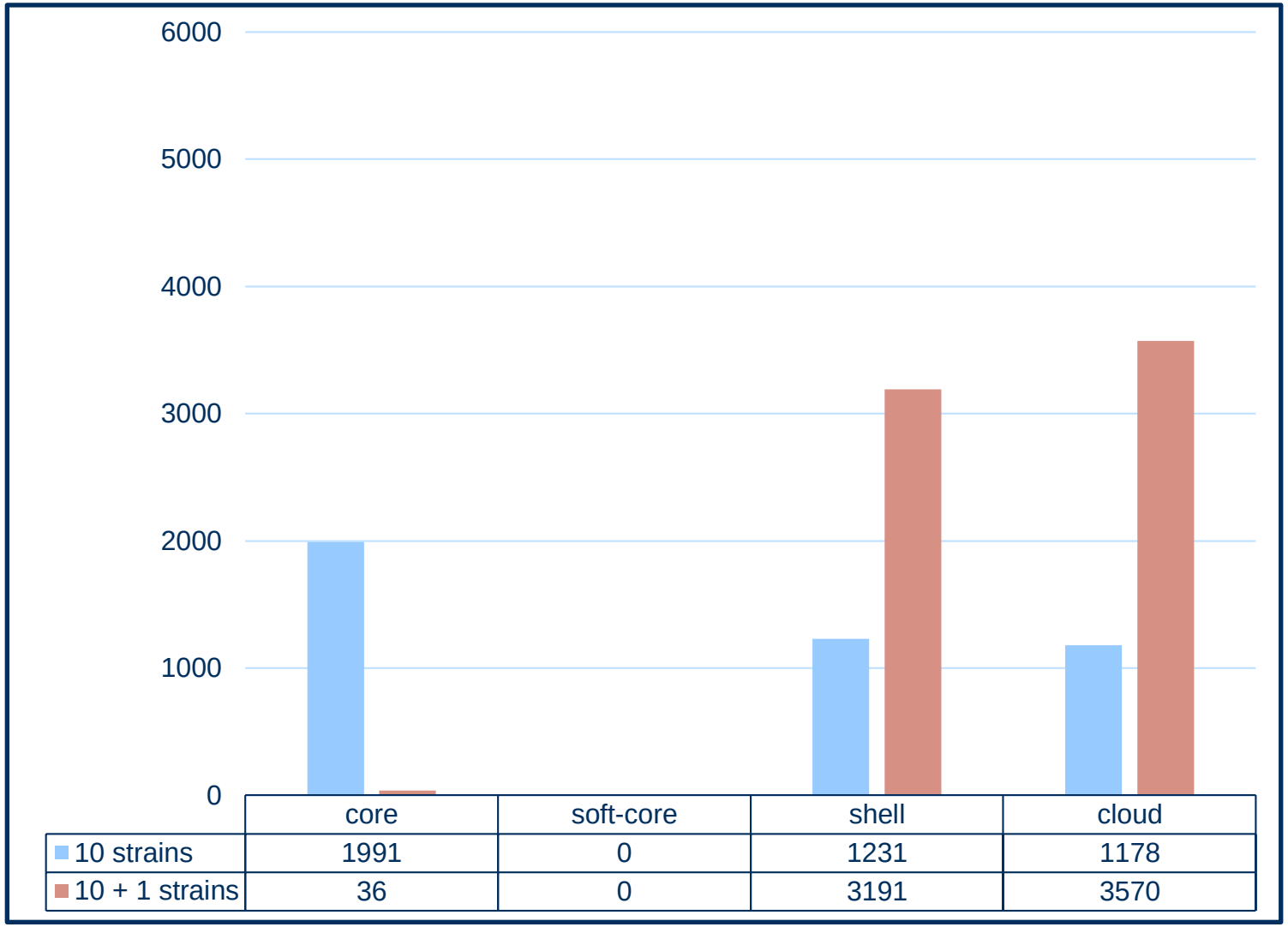


Sup. Fig. 13: A flowchart of the steps in the application.

The problem of interspecies pan-genomes
Staphylococcus aureus and *Staphylococcus saprophyticus*

1	>AP017922.1 Staphylococcus aureus DNA, strain: JP080
2	>CP013231.1 Staphylococcus aureus strain UTSW MRSA 55
3	>CP038021.1 Staphylococcus aureus strain 04-002
4	>CP038268.1 Staphylococcus aureus strain O55 isolate B118
5	>CP038819.1 Staphylococcus aureus strain O82
6	>CP039848.1 Staphylococcus aureus strain 2030RH1
7	>CP040623.1 Staphylococcus aureus strain D592-HR
8	>CP040801.1 Staphylococcus aureus strain S15
9	>LN626917.1 Staphylococcus aureus strain ILRI_Eymole1/1
10	>NC_002951.2 Staphylococcus aureus strain COL
11	>AP008934.1 Staphylococcus saprophyticus strain ATCC 15305 DNA

The problem of interspecies pan-genomes
Staphylococcus aureus and *Staphylococcus saprophyticus*

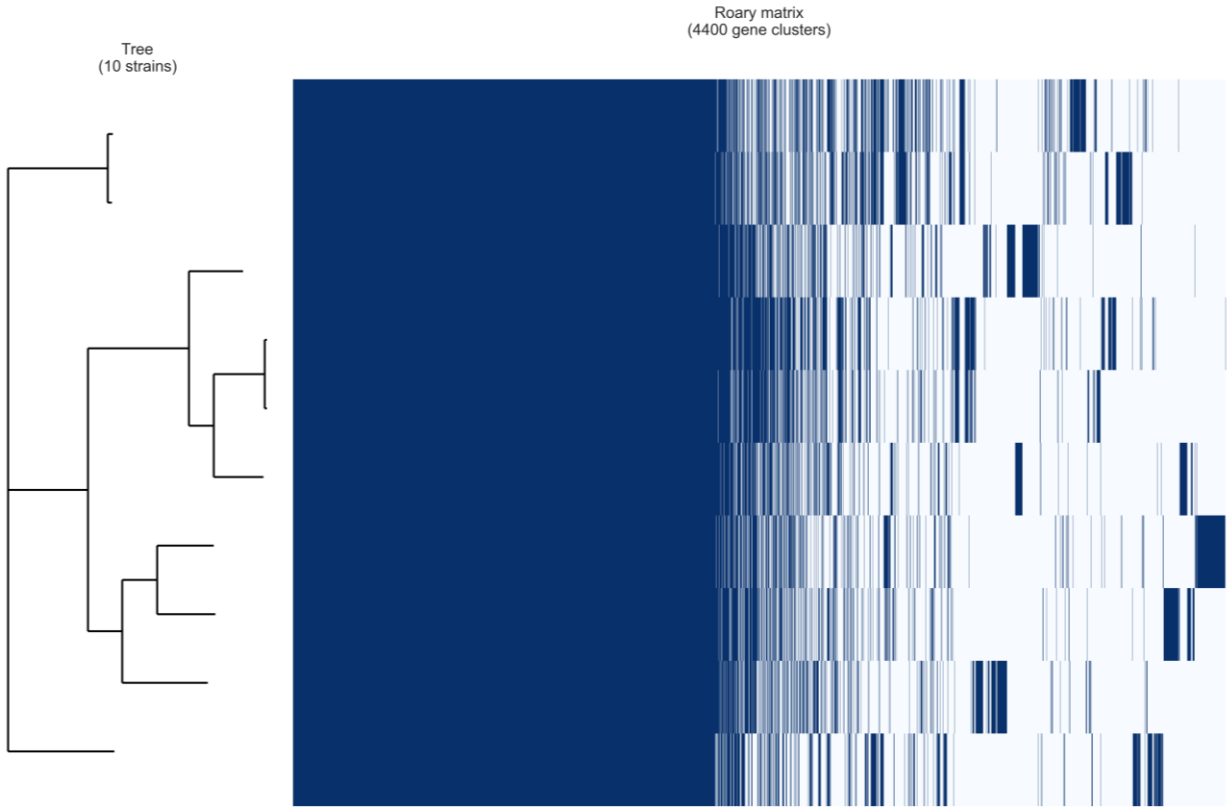


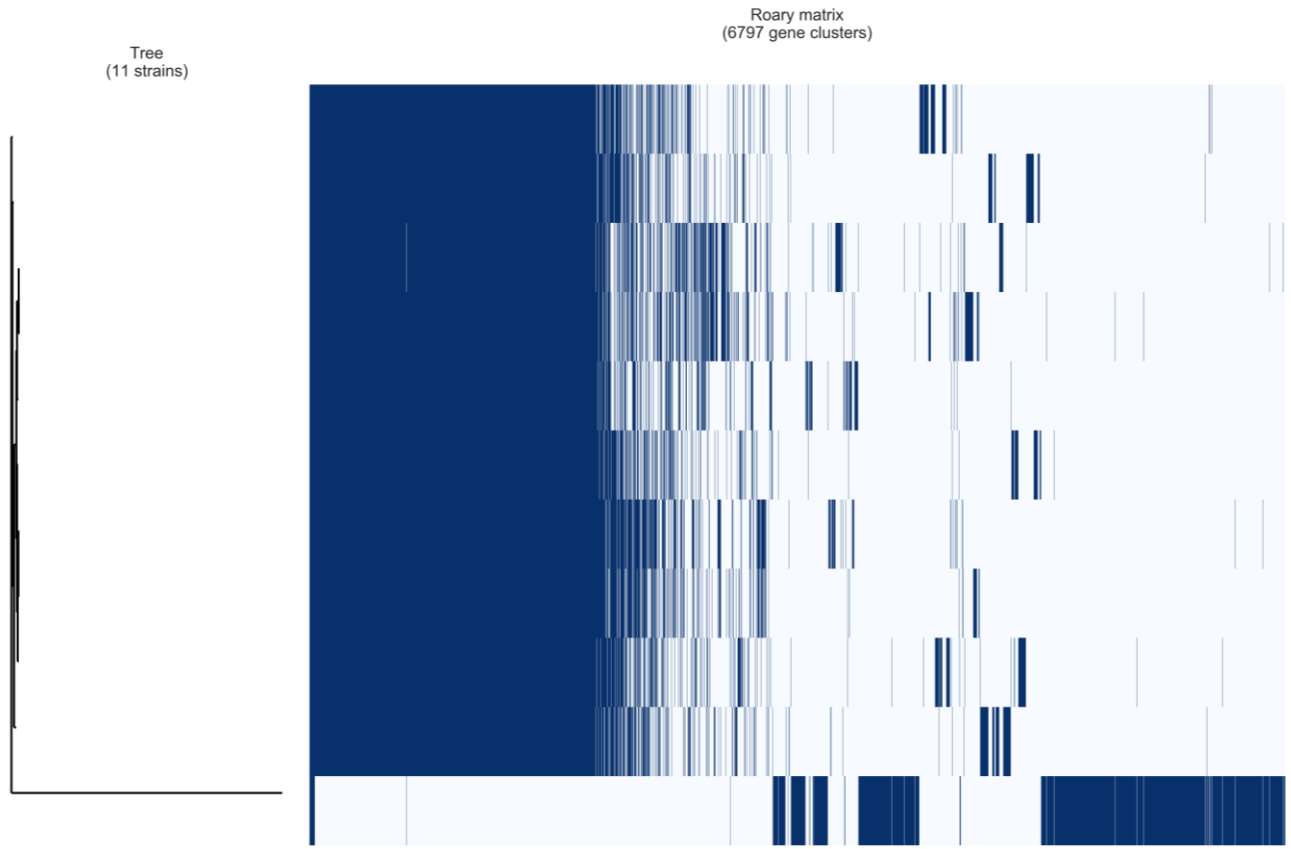
	core	soft-core	shell	cloud
10 strains	1991	0	1231	1178
10 + 1 strains	36	0	3191	3570

Core: 99% – 100%; soft-core: 96% – 98%; shell: 15% – 95%; cloud: < 15%

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

The problem of interspecies pan-genomes
Staphylococcus aureus and *Staphylococcus saprophyticus*





FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA