LECTURE SERIES

Klebsiella pneumoniae
Genomic Epidemiology and Antimicrobial Resistance

Bacterial strain taxonomy using LIN codes

Sylvain Brisse,
Institut Pasteur, Paris, France

Bacterial strain taxonomy using LIN codes

Sylvain Brisse

Director,
Biodiversity and Epidemiology of Bacterial Pathogens Unit
Natl. Ref. Center for Whooping Cough
Natl. Ref. Center for diphtheria
BIGSdb-Pasteur genomic libraries and strain taxonomy platform
Biological Resource Center of Institut Pasteur
Institut Pasteur, Paris, France

@sylvainbrisse

Sept 9th, 2025

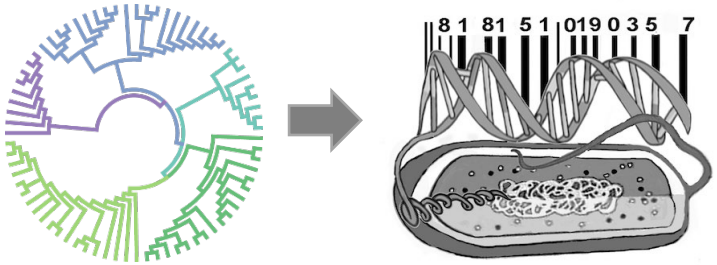
INSTITUT Pasteur

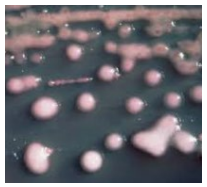
Biodiversity and Epidemiology of Bacterial Pathogens Unit

Emergence and evolution (of bacterial strains)

- Multidrug resistance
- Vaccine-escape
- Epidemiological surveillance
- One Health
- Links genotype-phenotype

Genomic taxonomies of strains



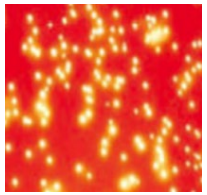


Klebsiella pneumoniae

Multidrug resistance

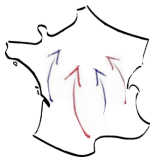
Corynebacterium diphtheriae

Multidrug resistance



Bordetella pertussis

Vaccine-driven evolution



Santé publique France

National Reference Center

EURL PUBLIC HEALTH DIPHTHERIA AND PERTUSSIS

Other related pathogens

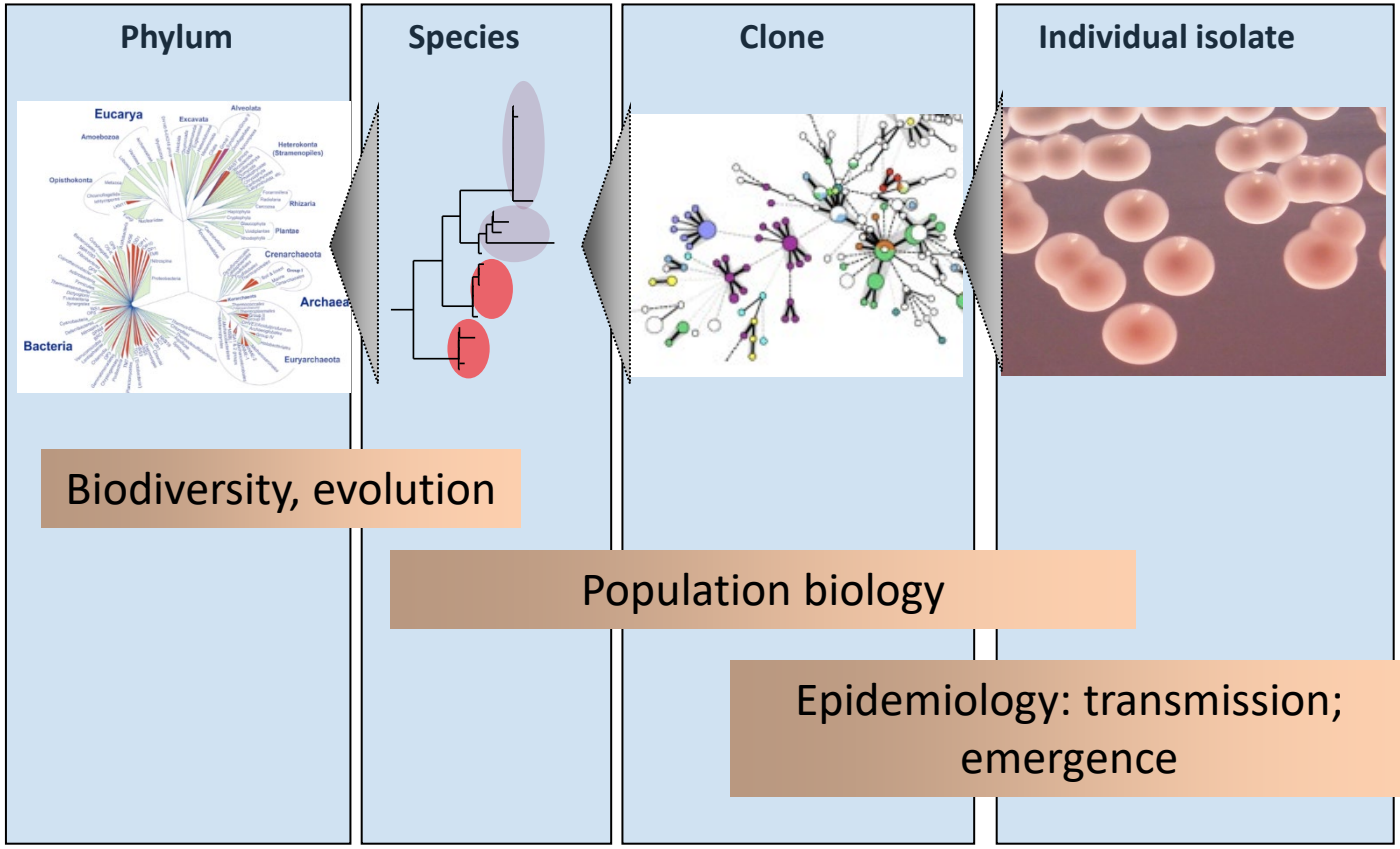
(of public health importance)

Intended Learning Objectives

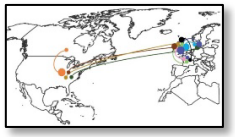
Specific objectives of this session:

- Comparative genomics approaches, including cgMLST
- What are LIN codes and how they work
- Links between LIN codes and the MLST nomenclature
- Integration into public strain taxonomy databases
- Demonstration and practical exercises with BIGSdb-Pasteur

Strain taxonomies: needs and applications



- Outbreak detection

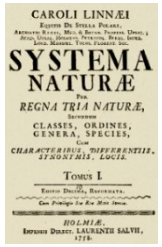


- Global spread

- Antimicrobial resistance

- Vaccine escape

Linnaean taxonomy ⟷ **Strain taxonomy**



?

Taxonomy

Classification
Delineating taxa

Nomenclature
Assigning names

Description
Properties of taxa

Identification
Determining to which taxon an individual belongs to
(Microbiological diagnostic)

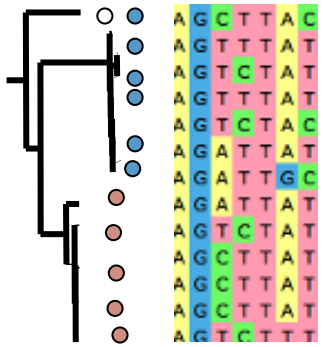
'Biological esperanto'

Taxonomy: requirements

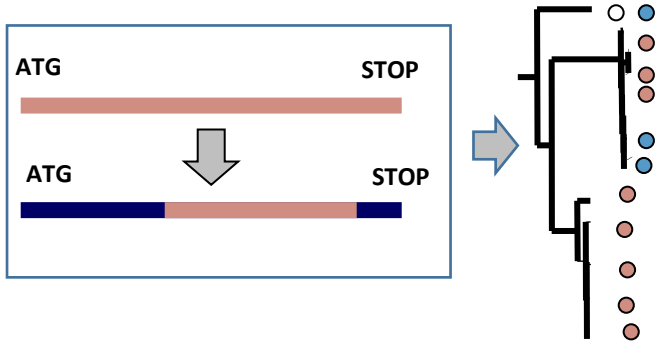
Taxonomy component	Needs
Classification	• Stability • Phylogenetic compatibility • Multiple scales (epidemiology, population biology)
Nomenclature	• Automatizable • Human readable • Backwards compatible (inheritability)
Identification	• Accurate • Accessible • Fast, user friendly

Bacterial evolution: drivers & implications

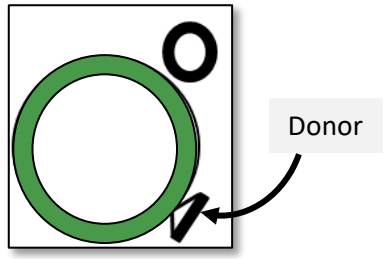
Mutation



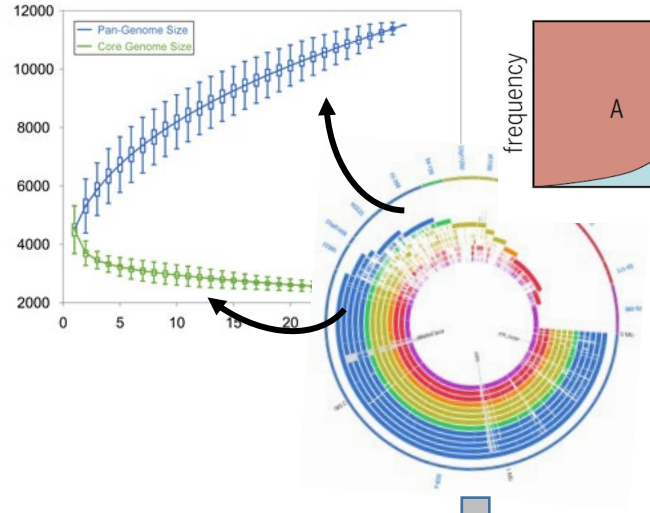
Recombination



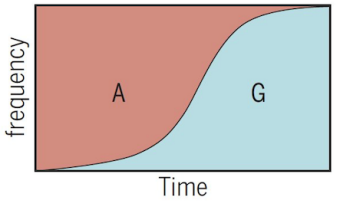
Gene transfer



Large gene pool



Selection



Core genes:

- Present in most strains, more neutral evolution, more vertical descent
- Maximize information content
- Minimize artefactual classifications

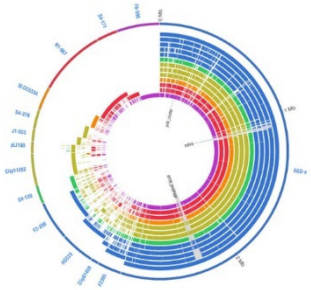
- Strain definition
- Epidemiological surveillance

Accessory genes:

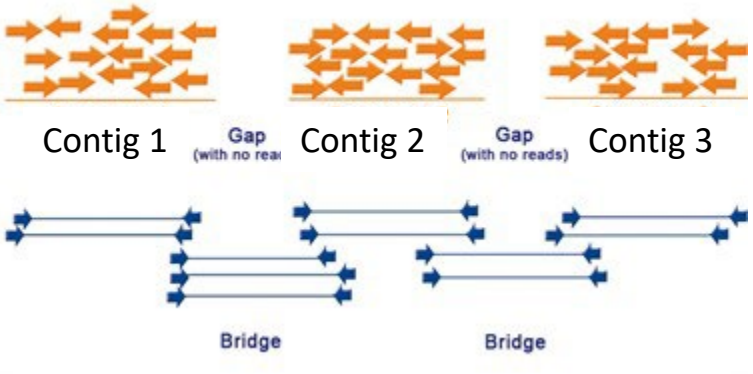
- Subject to strong selection & transfer

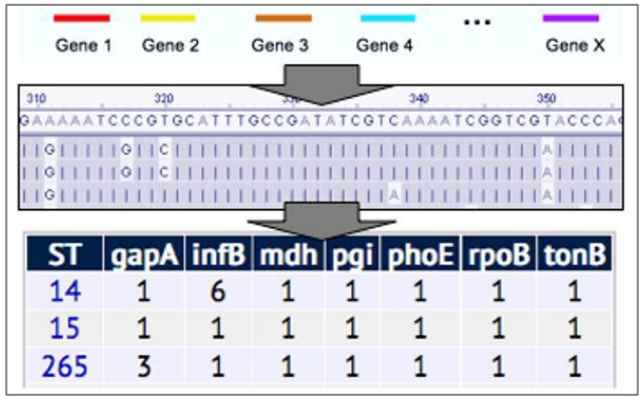
- Virulence elements
- Resistance elements
- Ecological adaptations

How do we compare bacterial genomes?
Genomic data analysis strategies



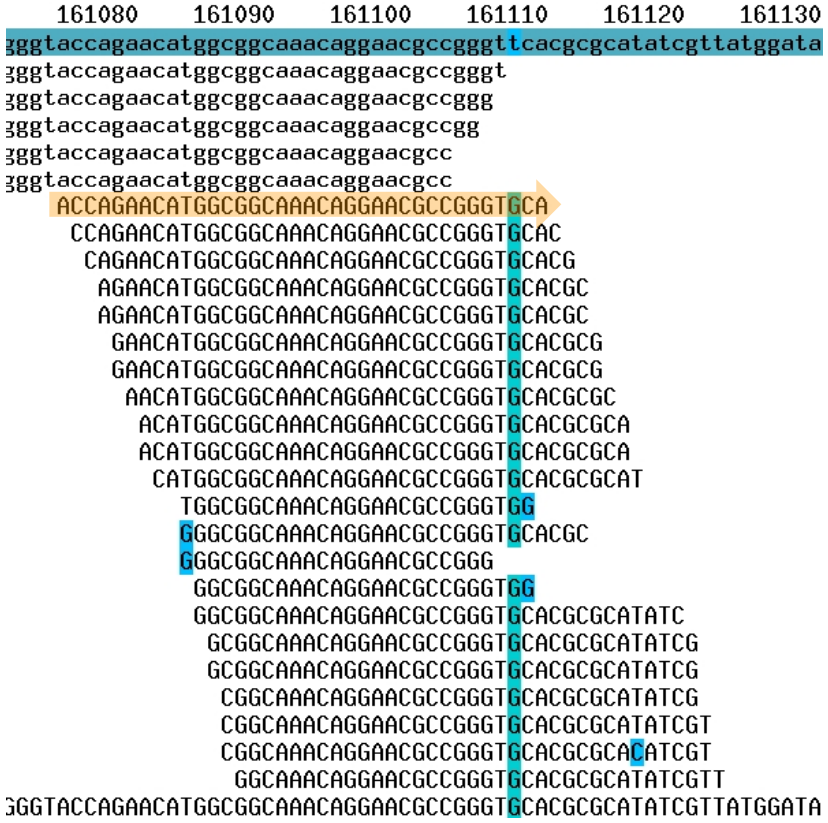
De-novo assembly



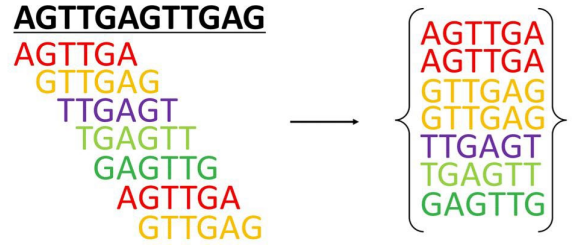


ST	gapA	infB	mdh	pgi	phoE	rpoB	tonB
14	1	6	1	1	1	1	1
15	1	1	1	1	1	1	1
265	3	1	1	1	1	1	1

Mapping



K-mers



- Count & compare
- Derive genetic distance (e.g., MASH distance)

Gene-by-gene approaches (MLST)

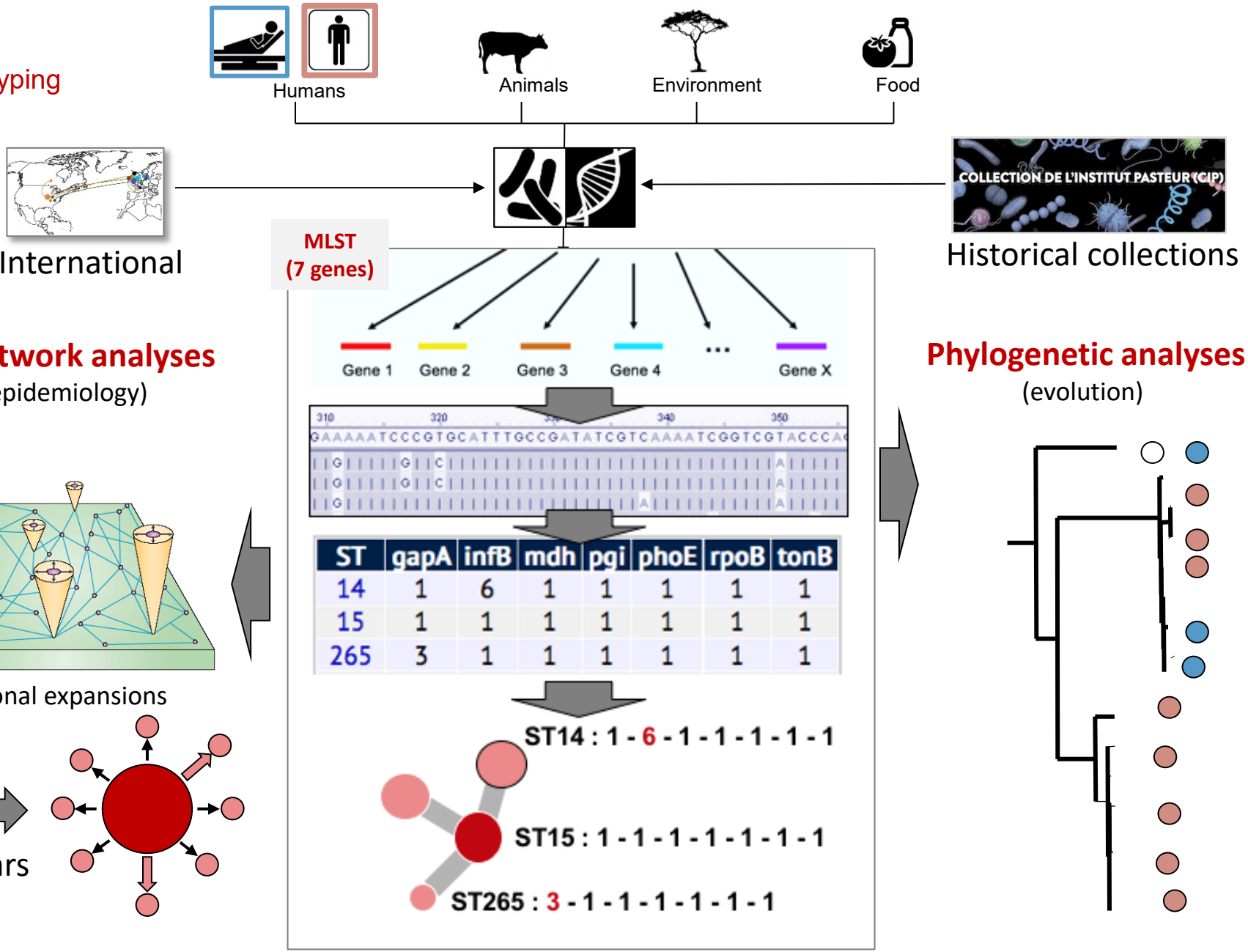
Single Nucleotide polymorphisms (SNPs)

K-mer approaches *(e.g.*, PopPuNK)

MLST: gene-by-gene, standardized, portable approach of strain diversity studies

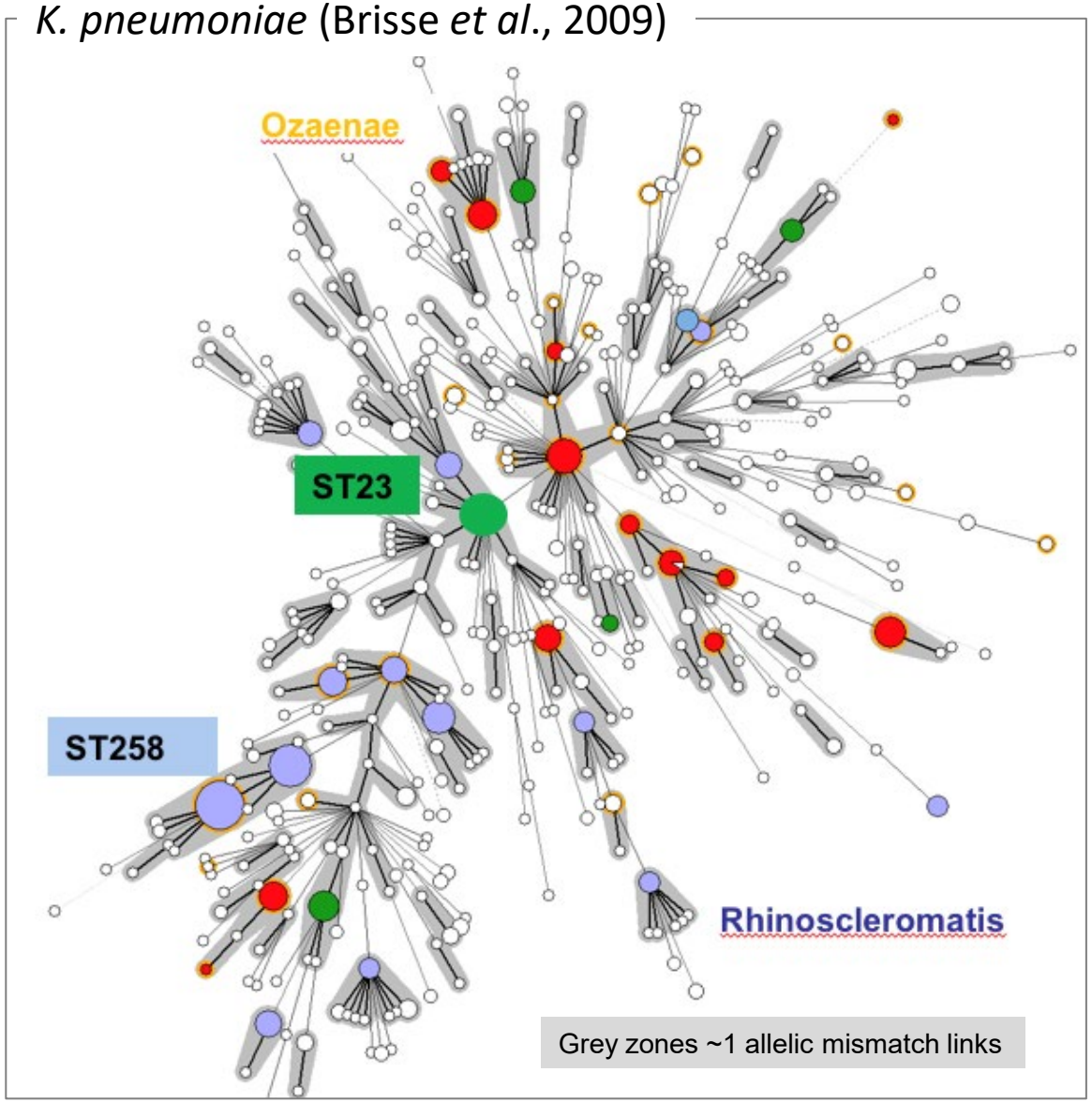
MLST:
Multilocus Sequence Typing

(Maiden *et al.*, 1998)



7-gene MLST : lack of resolution

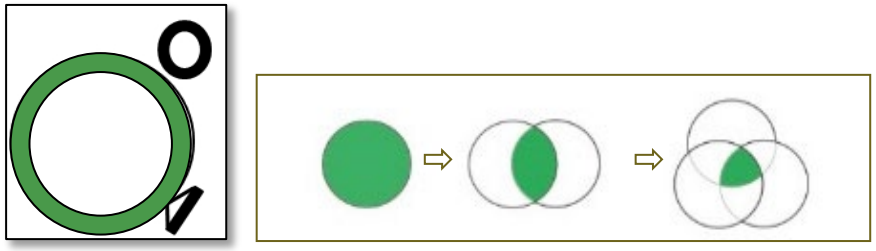
K. pneumoniae (Brisse *et al.*, 2009)



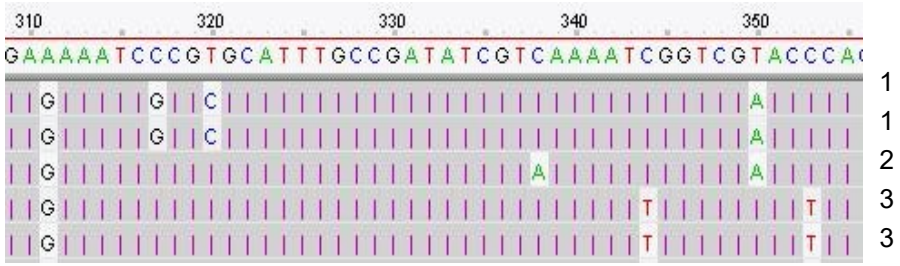
- 7 genes ~ 3500 nucleotides
- ~ 1/1000th of genome length
- Most STs are globally distributed
- STs can be > 100 years old
- ST do not inform on recent transmission

… but MLST is a common language

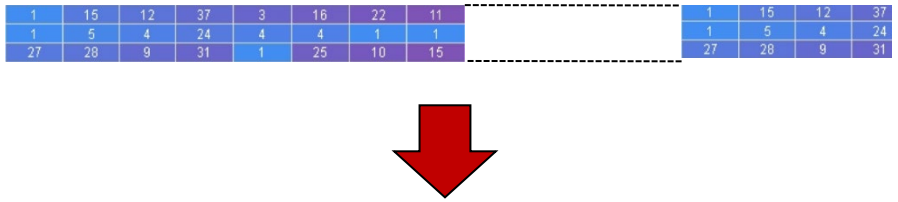
Core genome MLST



Step 1. Define core genome



Step 2. Define variation at core genes



Step 3. Define allelic profiles of genomes

1000-2000 gene loci

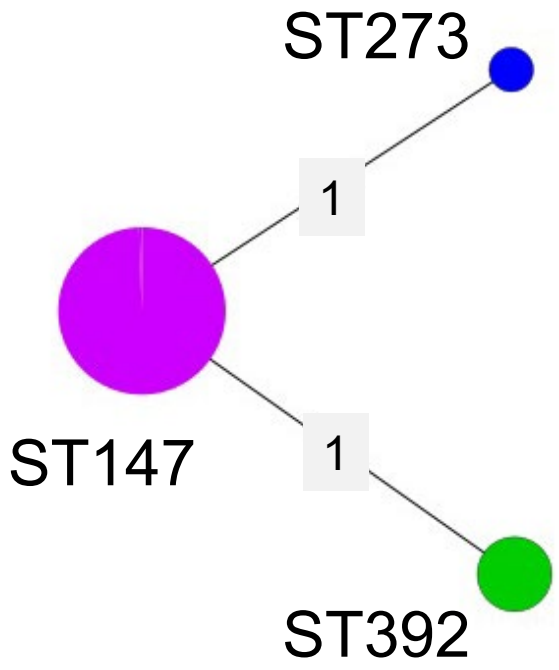
- Discrimination
- Classification precision

] are much improved compared to classical (7-gene) MLST

Klebsiella pneumoniae clonal group 147

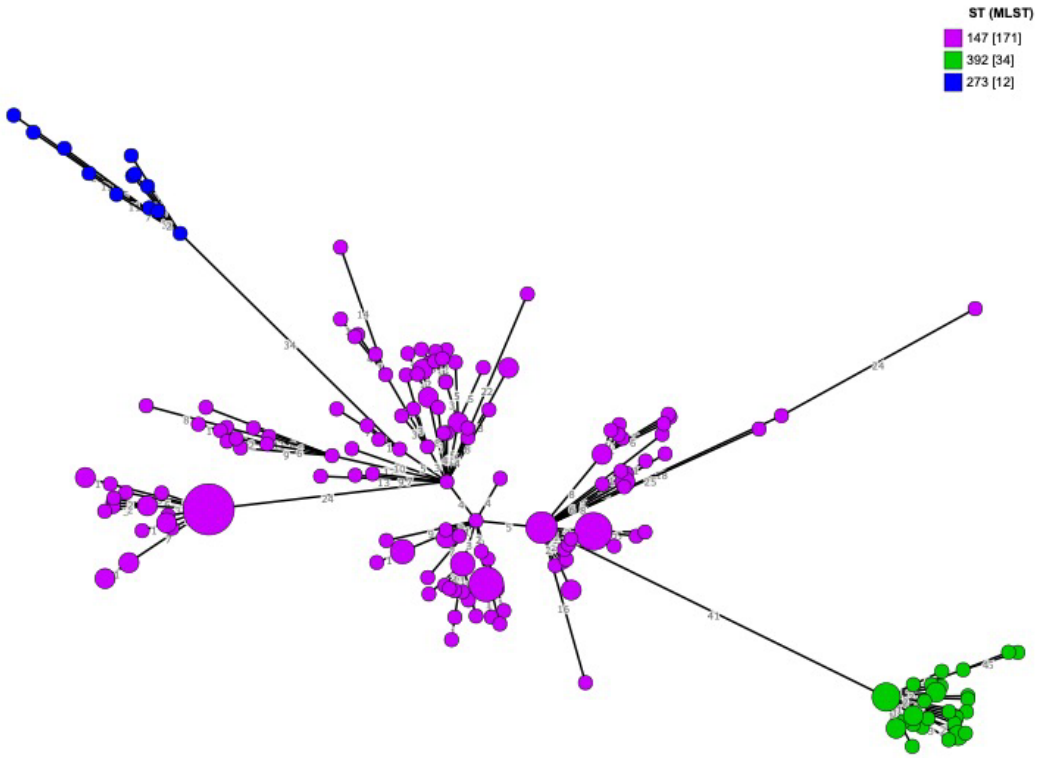
MStree based on MLST (7 genes)

(Diancourt et al., J Clin Microbiol 2005)



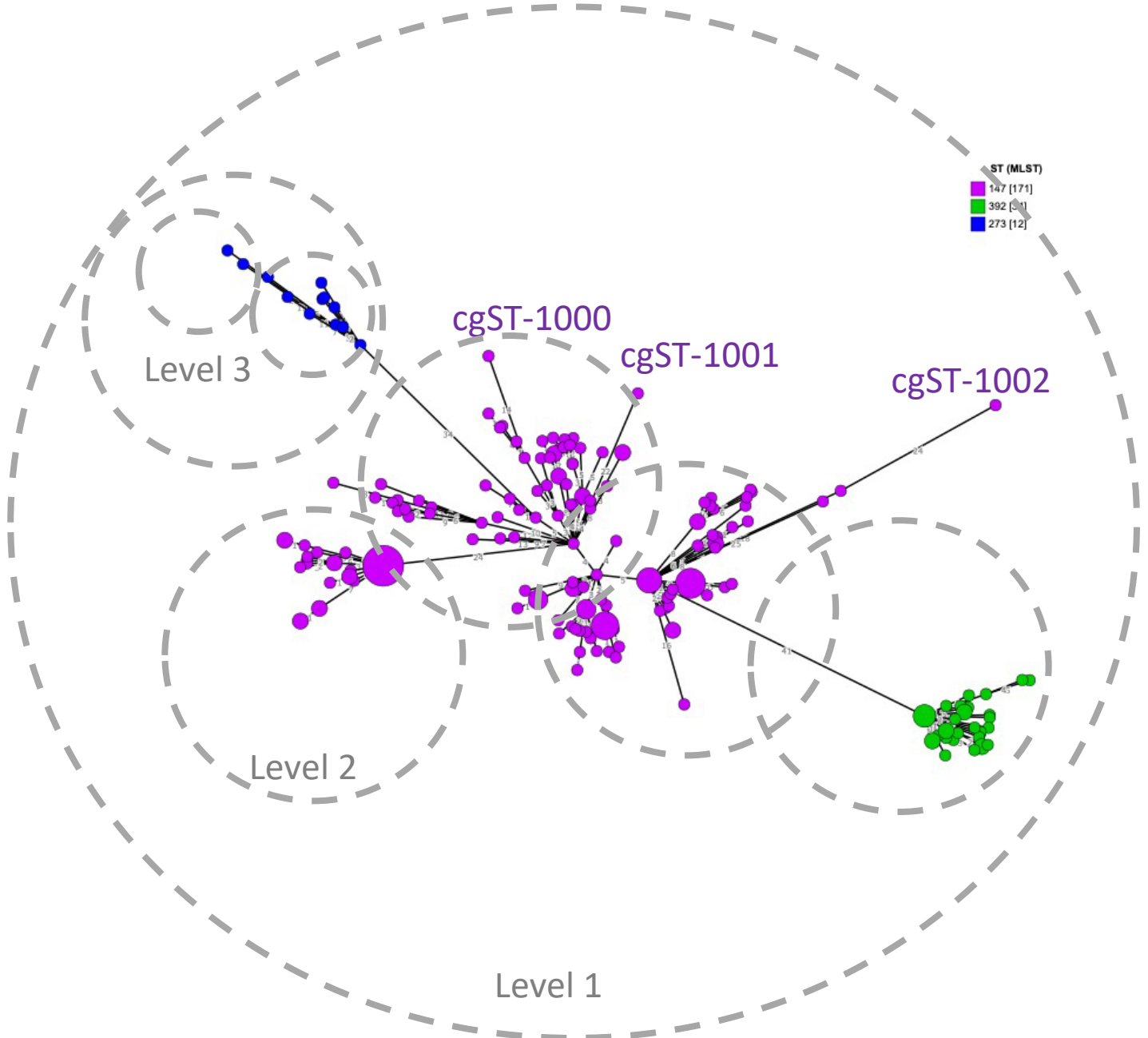
MStree based on cgMLST (629 genes)

(Hennart *et al*. Mol Biol Evol 2022)



MStree: minimum spanning tree

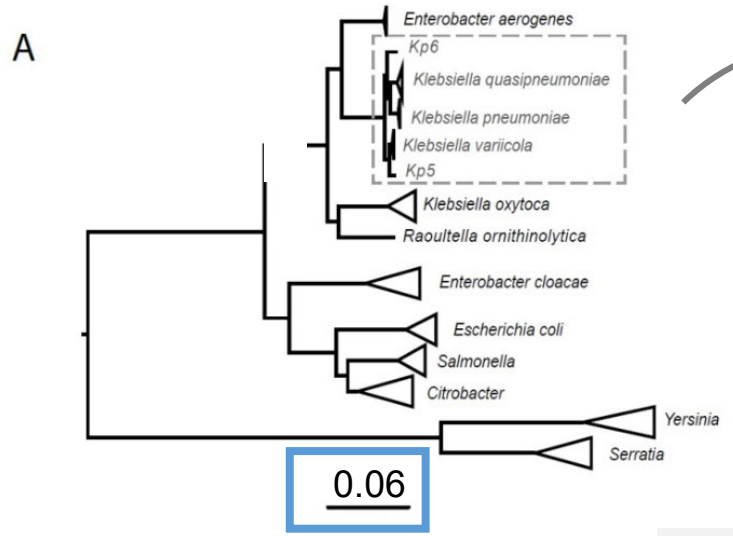
How can we classify bacteria using cgMLST profiles?



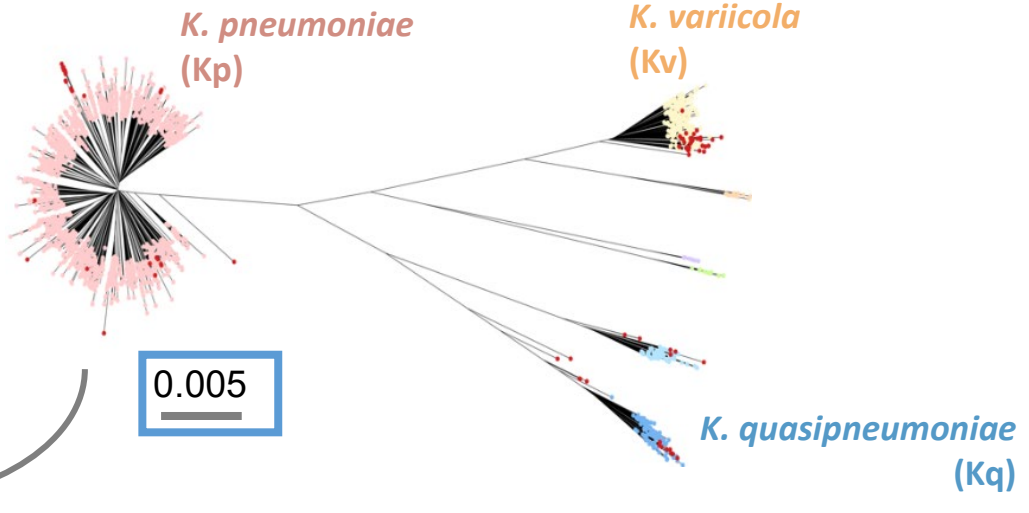
- **Problem**: cgSTs are too discriminatory for meaningful epidemiological associations
- **Solution**: group cgSTs based on their profiles' similarity levels

Need to choose relevant phylogenetic depths

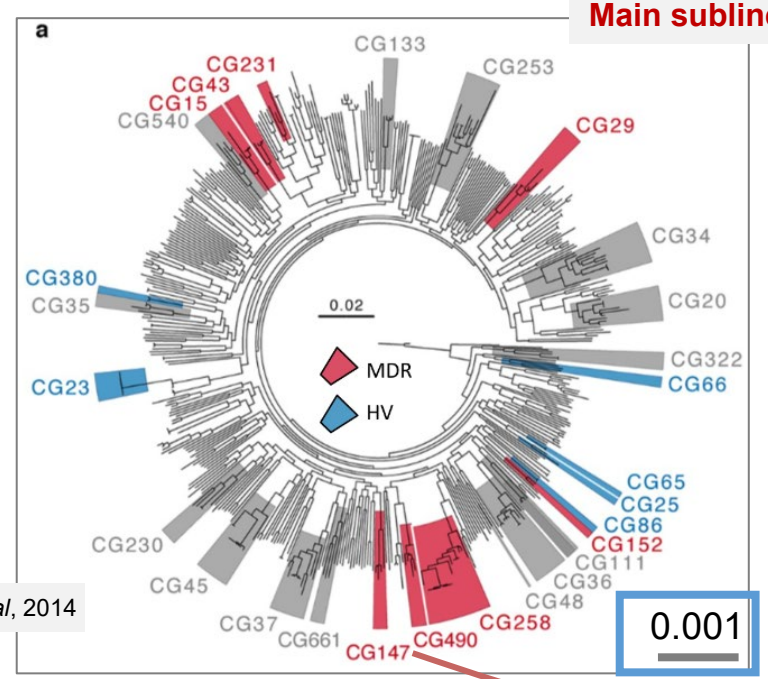
The multiple levels of phylogenetic structure



KpSC

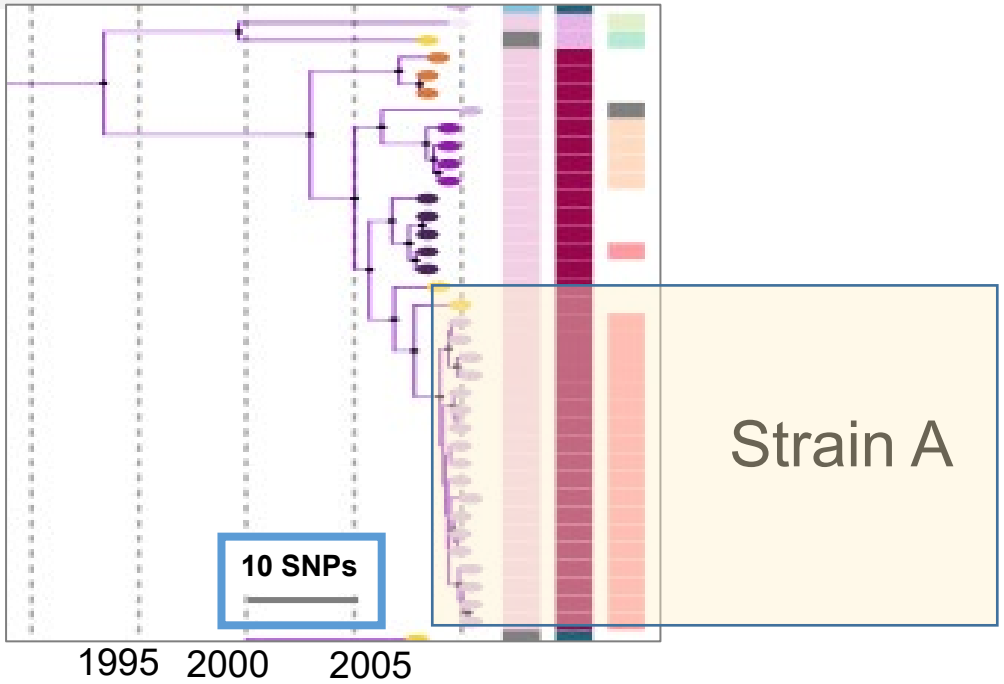


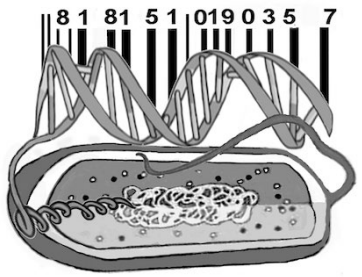
Main sublineages



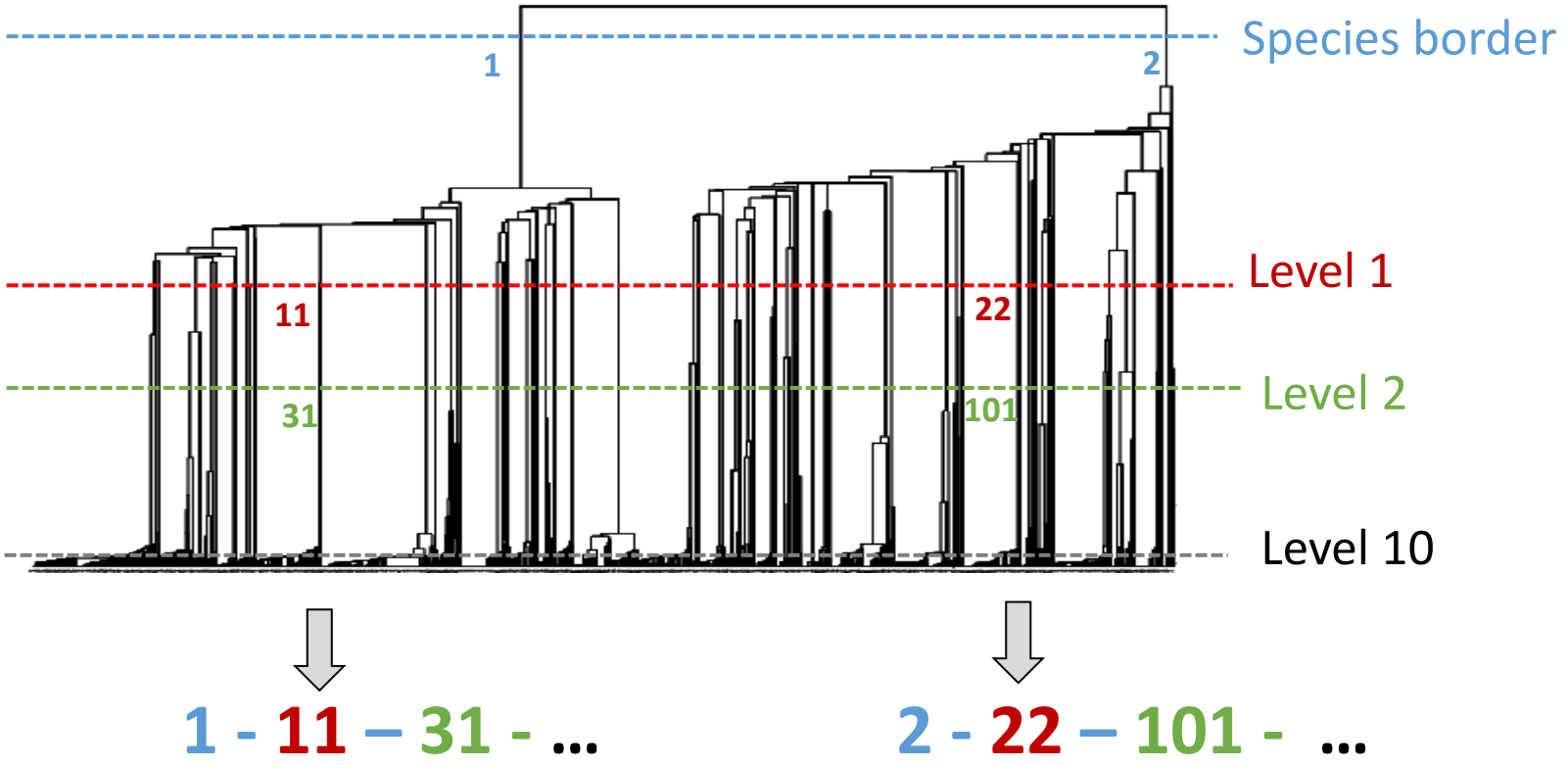
Holt *et al*, 2014

Emerging 'strains'



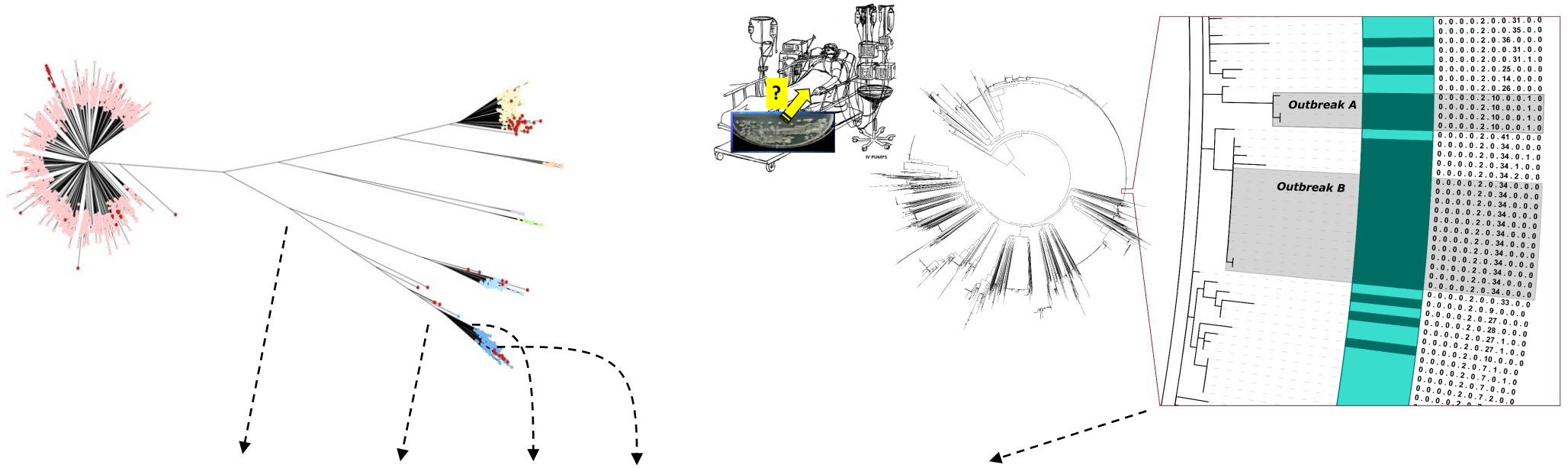


Multilevel classification



➡ Unique barcode for each strain, capturing group appartenance at various phylogenetic depths

10-level barcoding of *K. pneumoniae* Species Complex, from species to epidemiological tracking



	Species	Subspecies	Sublineage	Clonal group	Epidemiology					
Threshold	629	610	585	190	43	10	7	4	2	1

Number of accepted cgMLST mismatches (out of 629)

Core genome MLST Life Identification Number (LIN) codes for KpSC genomes

	Species	Subspecies	Sublineage	Clonal group	Epidemiology					
Threshold	629	610	585	190	43	10	7	4	2	1

Genome sequences → cgMLST profiles → cgST

	Bin number:	1	2	3	4	5	6	7	8	9	10
	Max. allelic similarity*:	19	44	439	586	619	622	625	627	628	629
	Min. allelic difference*:	610	585	190	43	10	7	4	2	1	0
		Bins left thresholds:									
	Closest genome (similarity %)	0	3.02	6.99	69.79	93.16	98.41	98.88	99.36	99.68	99.84	100
Genome A	Initialization	0	0	0	0	0	0	0	0	0	0	
Genome B	**A** (3.50%)	0	1	0	0	0	0	0	0	0	0	

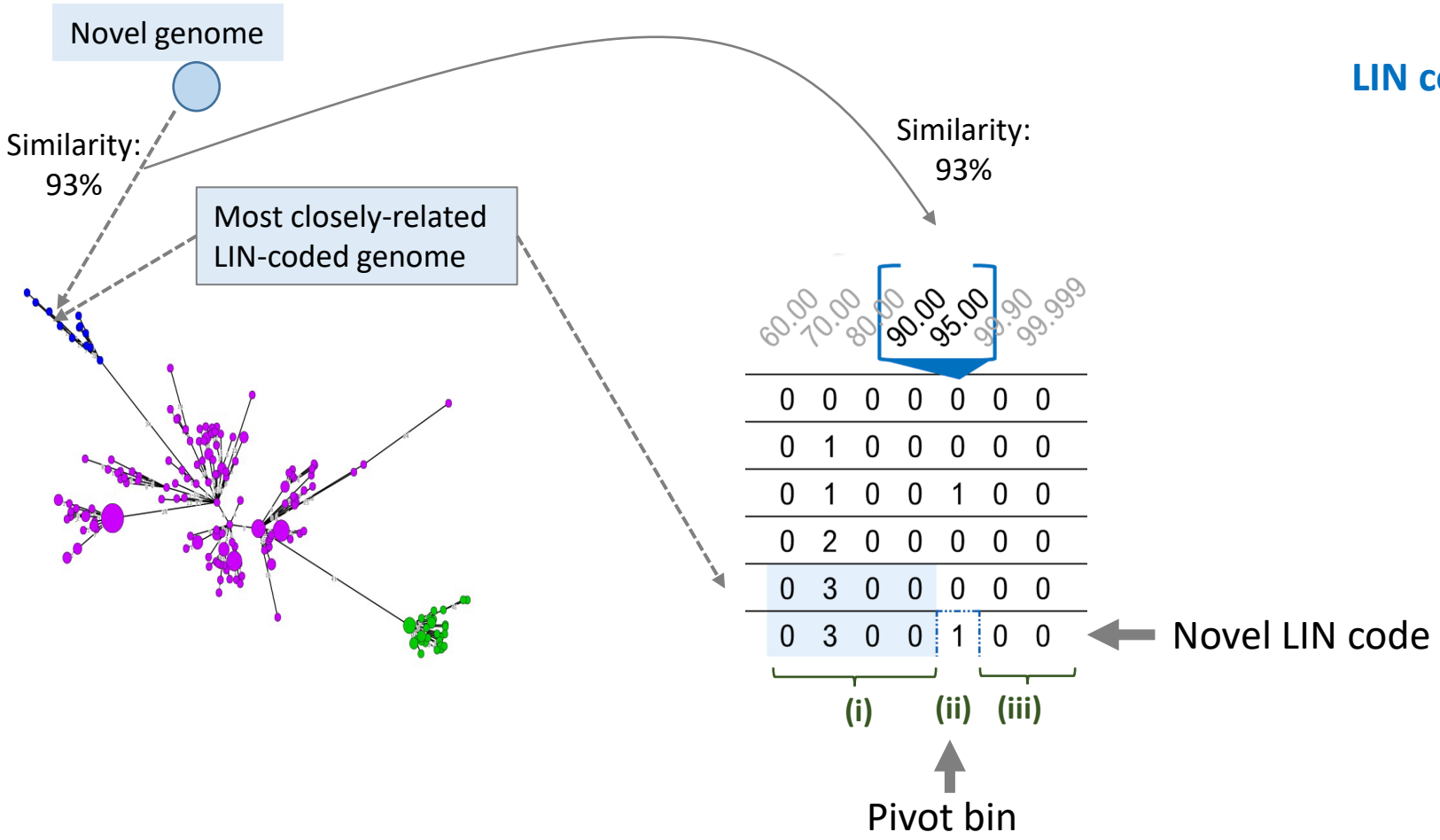
Palma, Hennart *et al.*, bioRxiv, 2024

Hennart *et al.,* Mol Biol Evol, 2022

LIN codes: encoding by proximity

1. Find closest genome in LIN database

2. Create LIN code based on closest neighbor's code, and similarity



LIN code creation steps:

1. Find pivot bin:

- the bin where the similarity value under consideration falls

2. Create novel LIN code:

(i) Prefix up to pivot bin
(ii) +1 in pivot bin
(iii) 0 in downstream bins

Core genome MLST Life Identification Number (LIN) codes: encoding process (KpSC example)

Genome sequences → cgMLST profiles → cgST

	Bin number:	1	2	3	4	5	6	7	8	9	10
	Max. allelic similarity*:	19	44	439	586	619	622	625	627	628	629
	Min. allelic difference*:	610	585	190	43	10	7	4	2	1	0
		Bins left thresholds:									
	Closest genome (similarity %)	0	3.02	6.99	69.79	93.16	98.41	98.88	99.36	99.68	99.84	100
Genome A	Initialization	0	0	0	0	0	0	0	0	0	0	
Genome B	**A** (3.50%)	0	1	0	0	0	0	0	0	0	0	
Genome C	**B** (99.0%)	0	1	0	0	0	0	1	0	0	0	
Genome D	**B** (7.00%)	0	1	1	0	0	0	0	0	0	0	
...	...	...										
Genome E	**D** (5.00%)	0	2	0	0	0	0	0	0	0	0	
Genome F	**E** (98.90%)	0	2	0	0	0	0	1	0	0	0	

Coding steps (genome E): (i) (ii) (iii)

* Right bin threshold, exclusive

Bin thresholds can be expressed in different ways

Similarity falls in pivot bin

Novel LIN code:

(i) Prefix up to pivot bin
(ii) +1 in pivot bin
(iii) 0 in downstream bins

Vinatzer *et al*. 2017 Antonie van Leeuwenhoek

Hennart *et al.*, Mol Biol Evol, 2022

How to read LIN codes

The shared prefix defines a similarity range: the range of their pivot bin

Genome sequences → cgMLST profiles → cgST

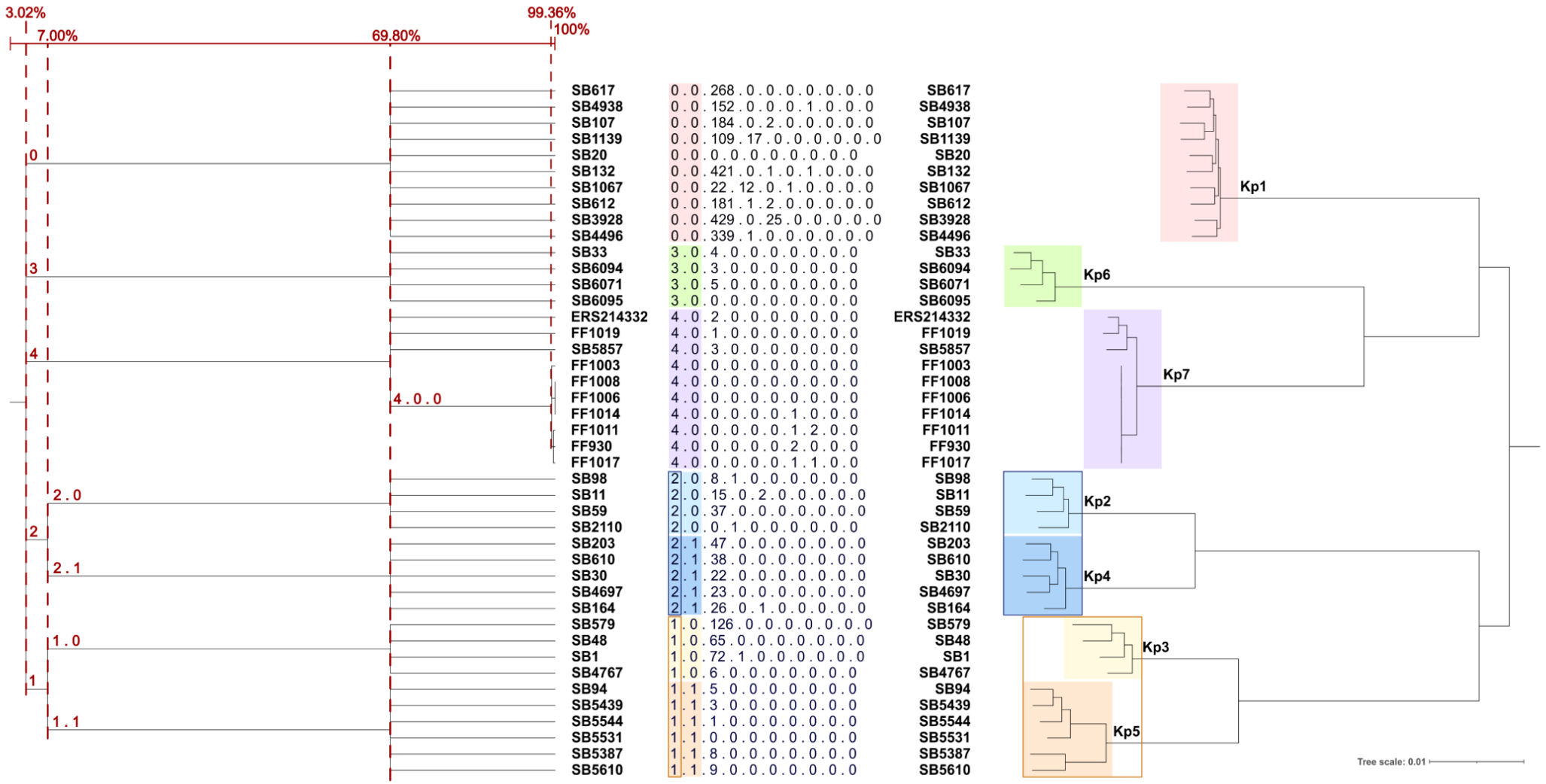
	Bin number:	1	2	3	4	5	6	7	8	9	10
	Max. allelic similarity*:	19	44	439	586	619	622	625	627	628	629
	Min. allelic difference*:	610	585	190	43	10	7	4	2	1	0
		Bins left thresholds:									
	Closest genome (similarity %)	0	3.02	6.99	69.79	93.16	98.41	98.88	99.36	99.68	99.84	100
Genome A	Initialization	0	0	0	0	0	0	0	0	0	0	
Genome B	**A** (3.50%)	0	1	0	0	0	0	0	0	0	0	
Genome C	**B** (99.0%)	0	1	0	0	0	0	1	0	0	0	
Genome D	**B** (7.00%)	0	1	1	0	0	0	0	0	0	0	
...	...	...										
Genome E	**D** (5.00%)	0	2	0	0	0	0	0	0	0	0	
Genome F	**E** (98.90%)	0	2	0	0	0	0	1	0	0	0	
	Coding steps (genome E):							(i)	(ii)		(iii)	

At least 6.99% similar, at most 69.79 (exclusive)

At least 98.88% similar, at most 99.36 (exclusive)

* Right bin threshold excluded

LIN codes prefixes are diagnostic markers



- Prefixes mark important phylogroups, sublineages, clonal groups and finer subdivisions
- Can be used to label such groups

LIN codes levels are nested from right to left

LIN prefix, size 2
0_0
1_0
1_1
2_0
2_1
3_0
4_0

LIN prefix, size 3
0_0_0
0_0_429
0_0_105
0_0_158
0_0_197
0_0_369
0_0_750

LIN prefix, size 4
0_0_105_6
0_0_105_0
0_0_105_2
0_0_105_11
0_0_105_1
0_0_105_29
0_0_105_7

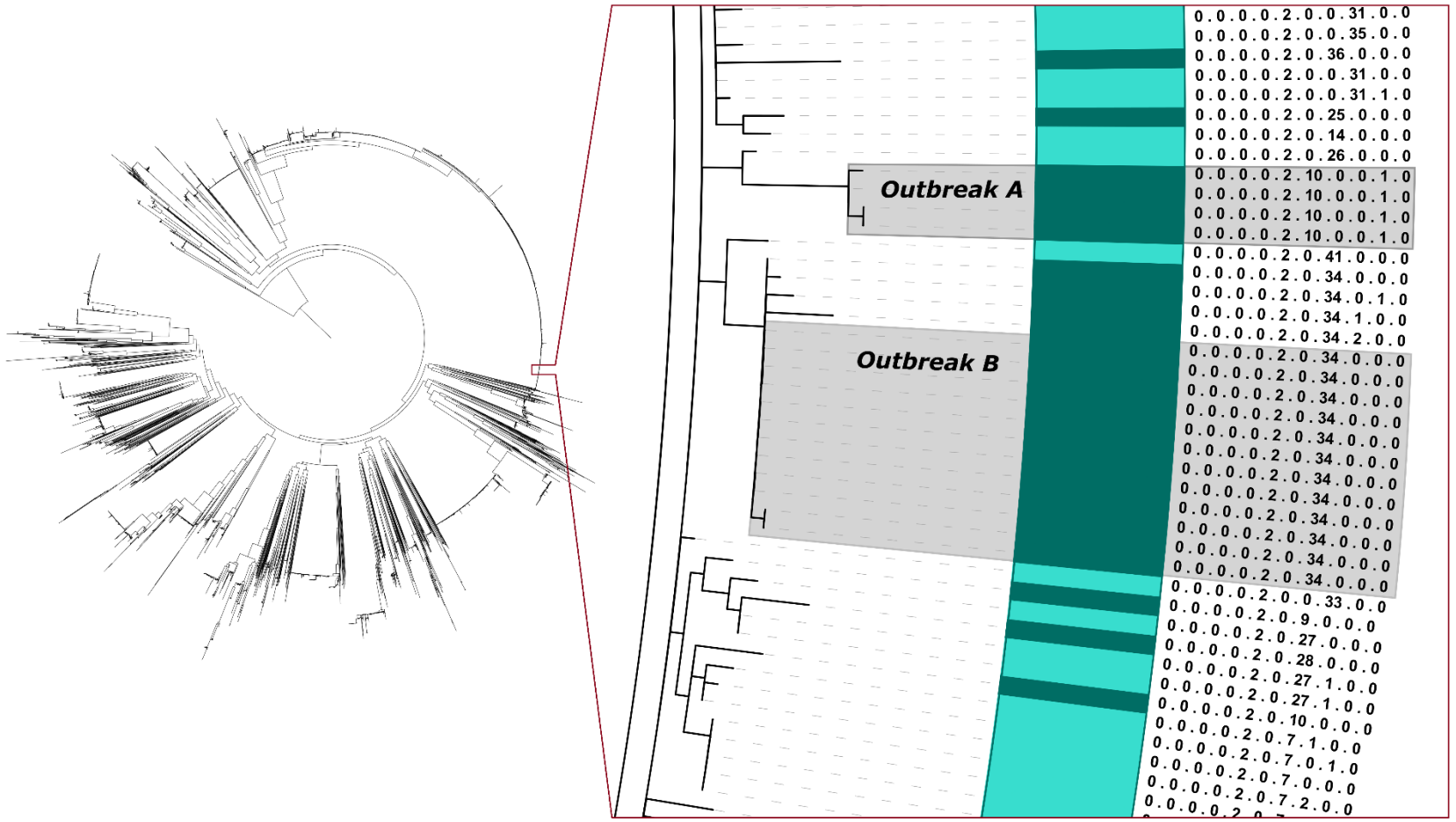
LIN prefix
0_0_1_**0**
0_0_2_**0**
0_0_105_**0**

0 has a different meaning in these three lines

(think of *K. pneumoniae* versus *Streptococcus pneumoniae*)

LIN codes are predominantly composed of small integers

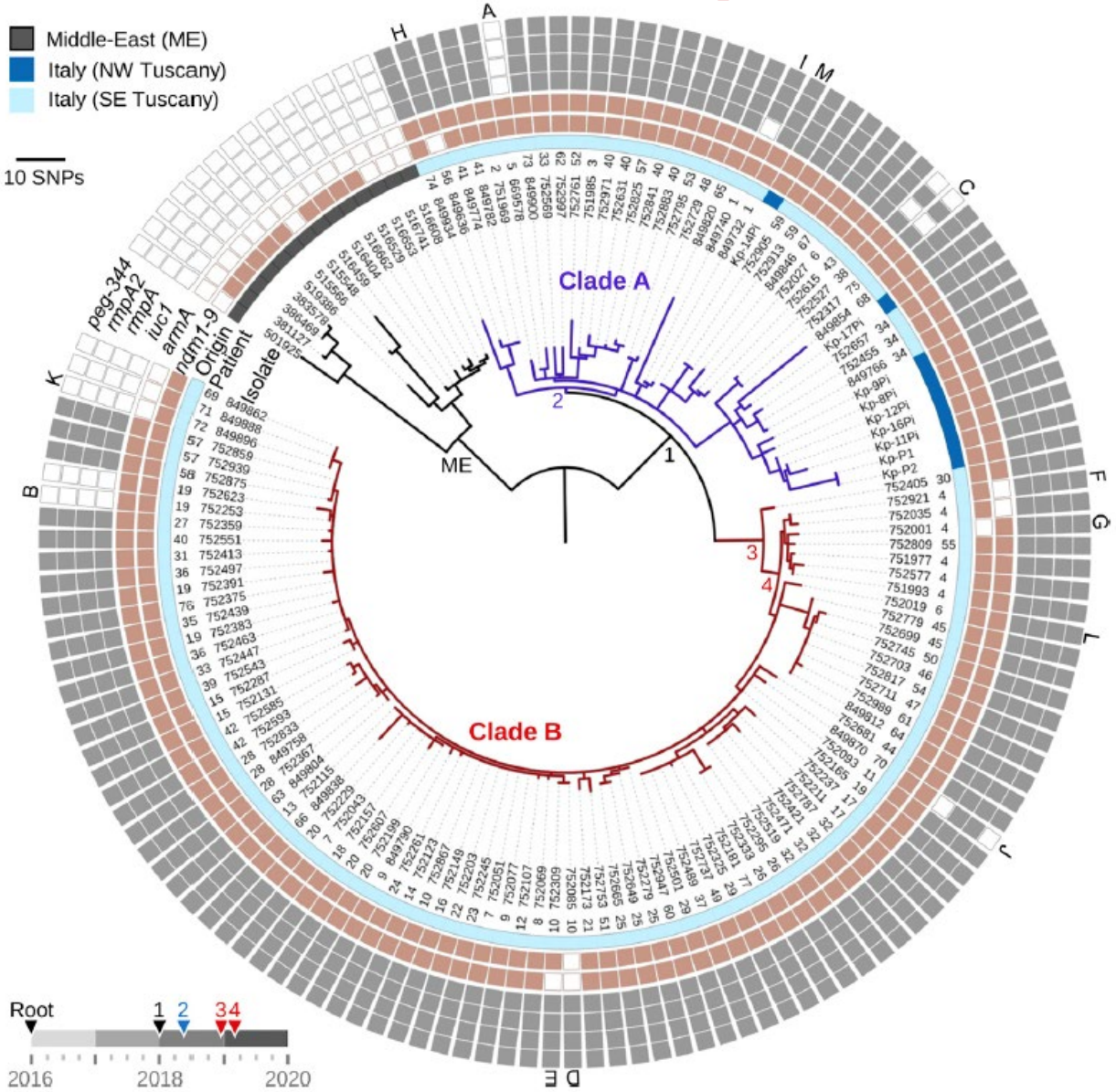
LIN codes can be used to track outbreak strains



Only the almost identical genomes may have same LIN code (although for *K. pneumoniae*, the 629-loci cgMLST scheme has somewhat restricted discrimination)

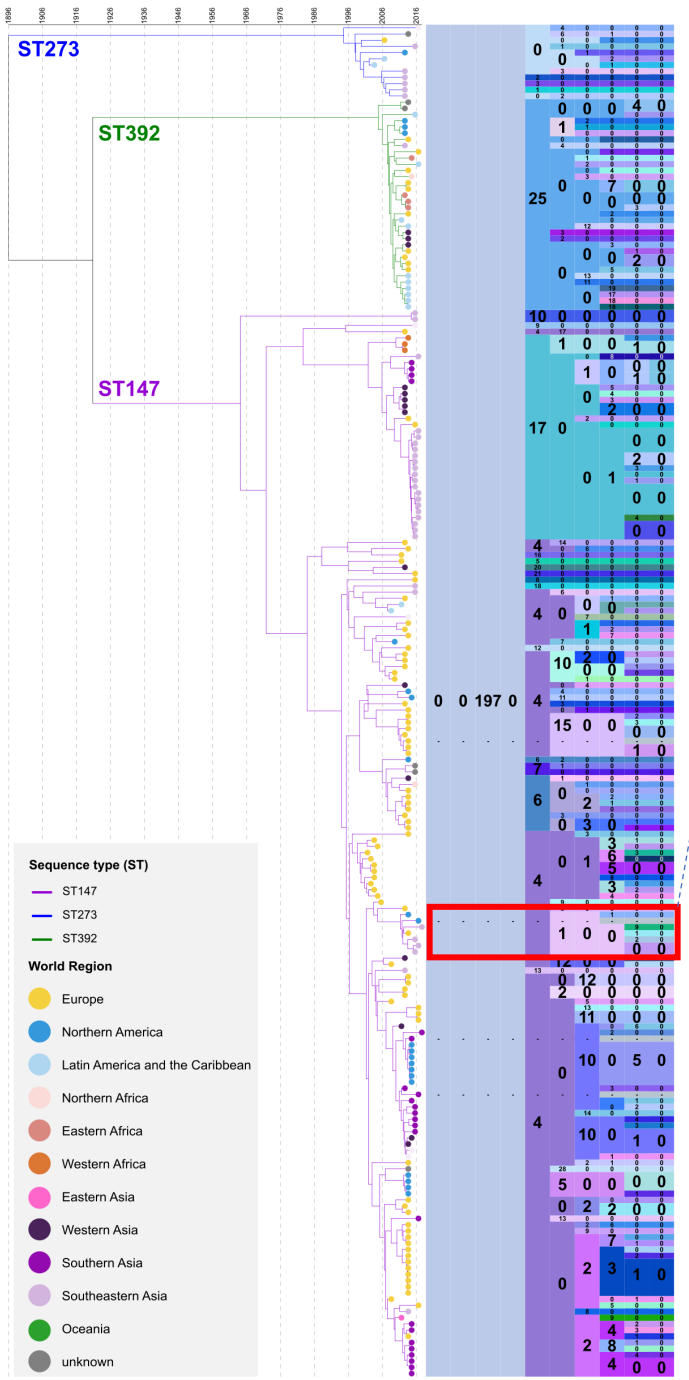
Anatomy of an extensively drug-resistant *Klebsiella pneumoniae* outbreak

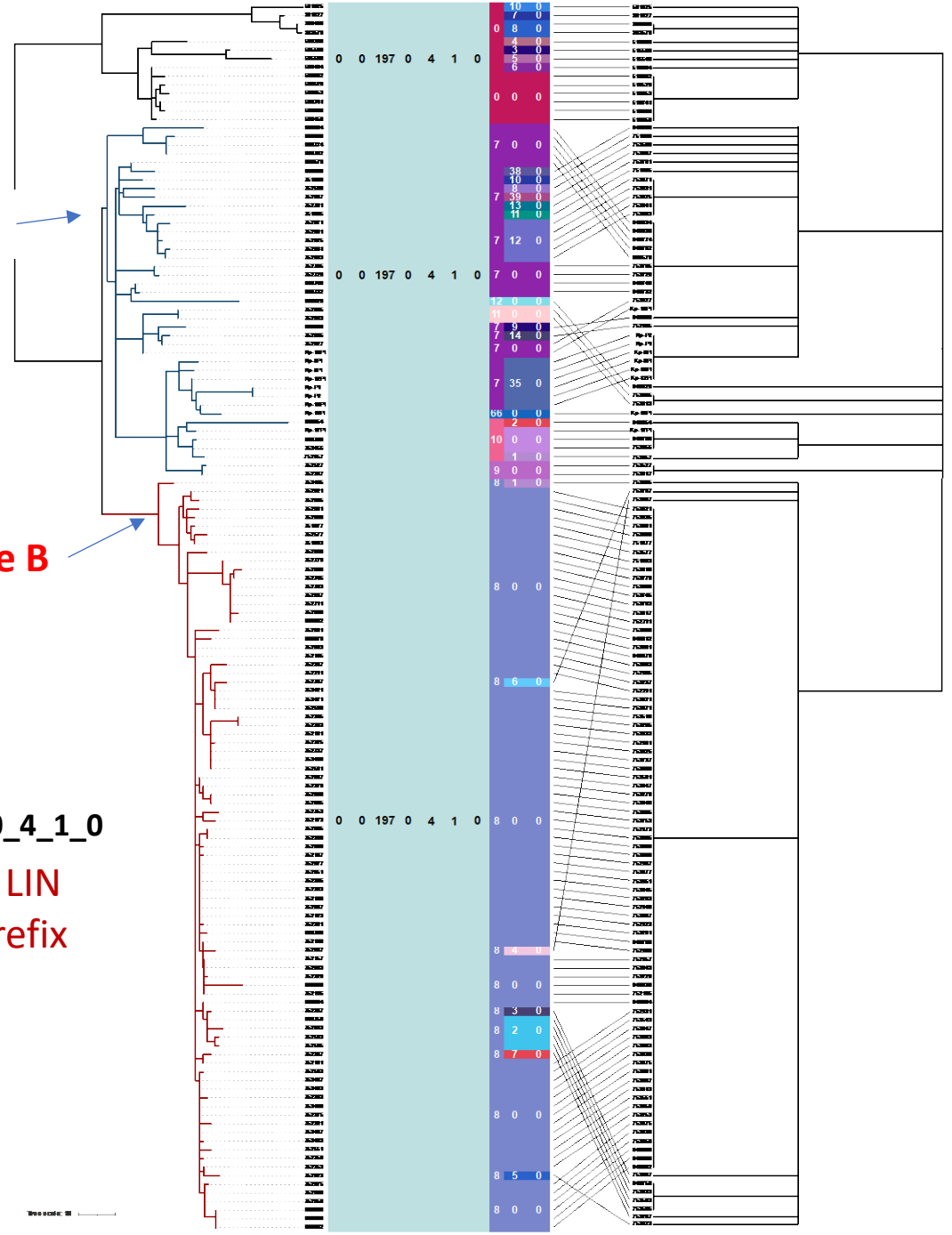
Martin *et al*., 2021, PNAS



- Protracted outbreak of NDM-1 ST147, Italy, 2018-2021
- Shared a recent ancestor with clinical isolates from the Middle East (>45 SNPs).
- Acquisition of convergent IncFIB/IncHIB–type plasmid carrying iuc1, rmpA, armA
- Increased siderophore production, but not hypervirulence
- Independent emergence of resistance to either fosfomycin (glpT), tigecycline (ramR), or colistin (mgrB)
- Clades A and B: 33 SNPs on average
- Clade A heterogeneous (11 SNPs avg from nearest neighbor); several hospitals
- Clade B homogeneous (2.4 SNPs avg from nearest neighbor); mostly single hospital

LIN codes of global SL147 isolates





0_0_197_0_4_1_0

Shared LIN code prefix

Clade A: multiple LIN codes

Clade B: mostly one LIN code

(~ 1 year sampling frame)

LIN codes of Italian SL147 outbreak

Taxonomy: requirements

Taxonomy component	Needs
Classification	• Stability • Phylogenetic compatibility • Multiple scales (epidemiology, population biology)
Nomenclature	• Automatizable • Human readable • Backwards compatible (inheritability)
Identification	• Accurate • Accessible • Fast, user friendly

LIN codes: saying goodbye politely to MLST

Sublineages and clonal groups nicknames ensure continuity of the main MLST identifiers within the LIN codes taxonomy

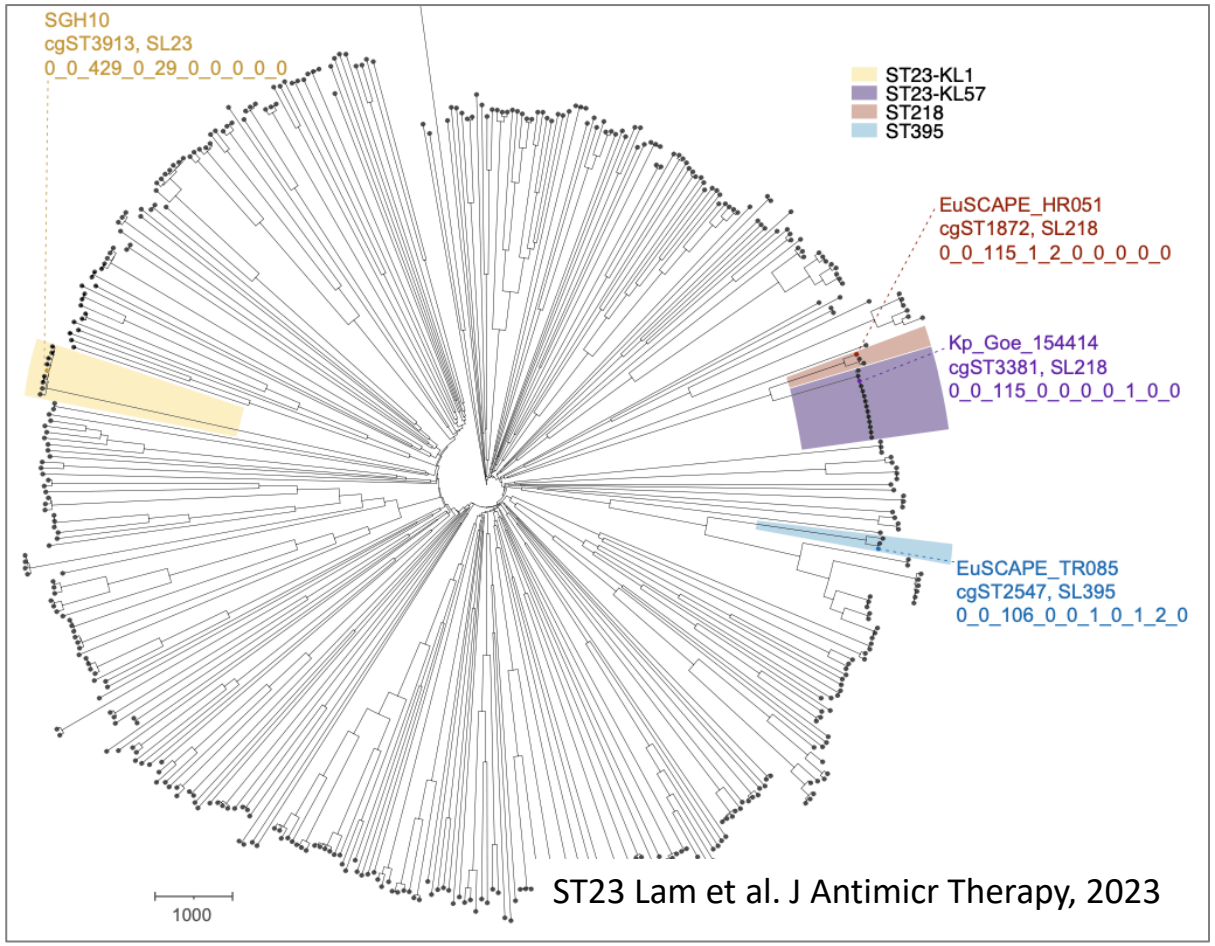
LIN prefix	Phylo-group
0_0	Kp1
1_0	Kp3
1_1	Kp5
2_0	Kp2
2_1	Kp4
3_0	Kp6
4_0	Kp7

LIN prefix	Main ST	Nickname
0_0_0	15	SL15
0_0_429	23	SL23
0_0_105	258	SL258
0_0_158	45	SL45
0_0_197	147	SL147
0_0_369	307	SL307
0_0_750	6589	SL10691

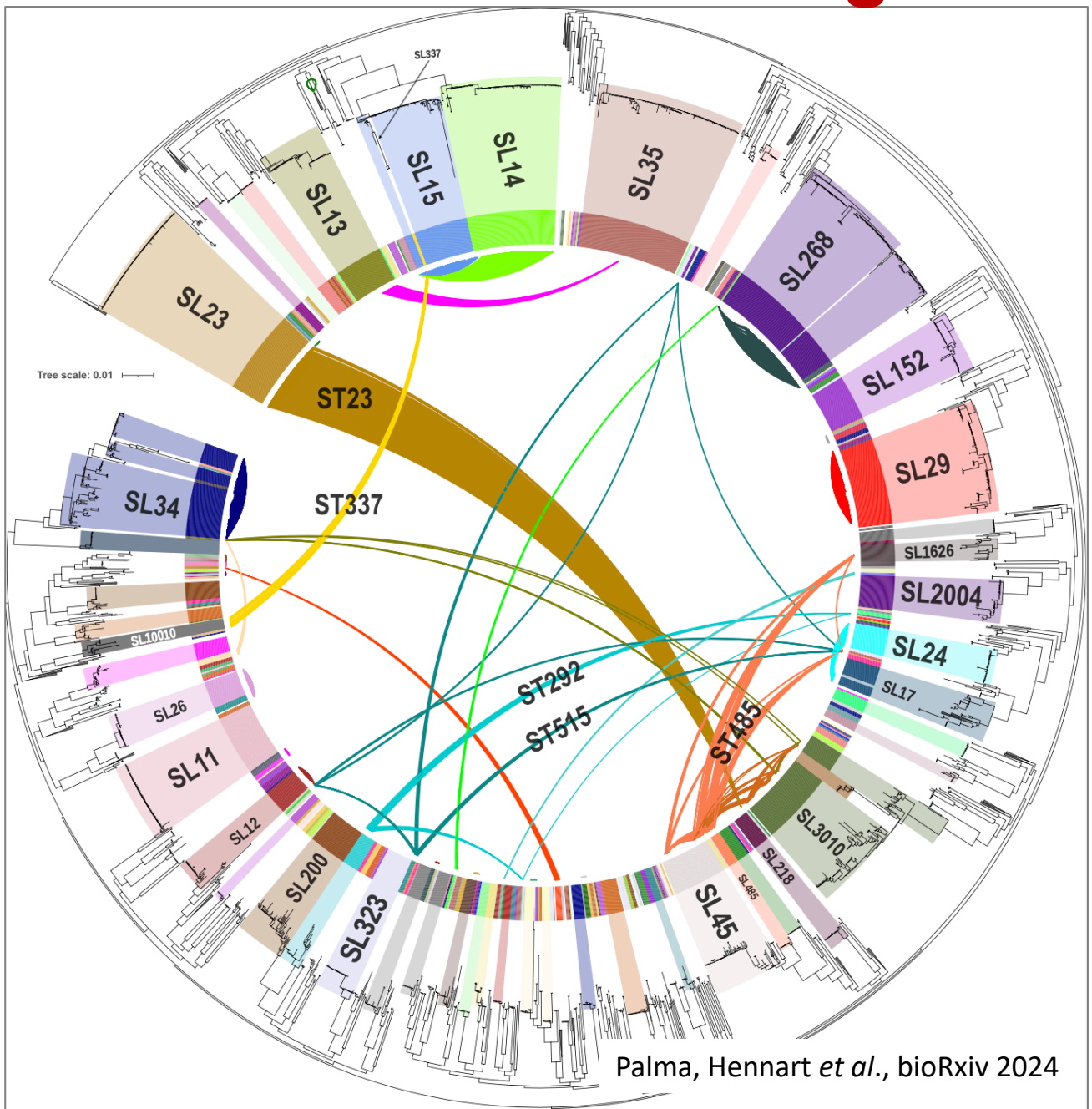
LIN prefix	Main ST	Nickname
0_0_105_6	258	CG258
0_0_105_0	340	CG340
0_0_105_2	11	CG11
0_0_105_11	11	CG3666
0_0_105_1	437	GC10268
0_0_105_29	11	CG12811
0_0_105_7	895	CG895

- A human-readable, backwards compatible (MLST) nomenclature
- Keep promoting MLST for Sanger sequencing-based labs & familiar system expansion

Some *Klebsiella* MLST STs can be misleading



ST23 Lam et al. J Antimicr Therapy, 2023



Palma, Hennart *et al.*, bioRxiv 2024

- Large-scale recombinations create distant sublineages
- Convergence of MLST profiles by recombination

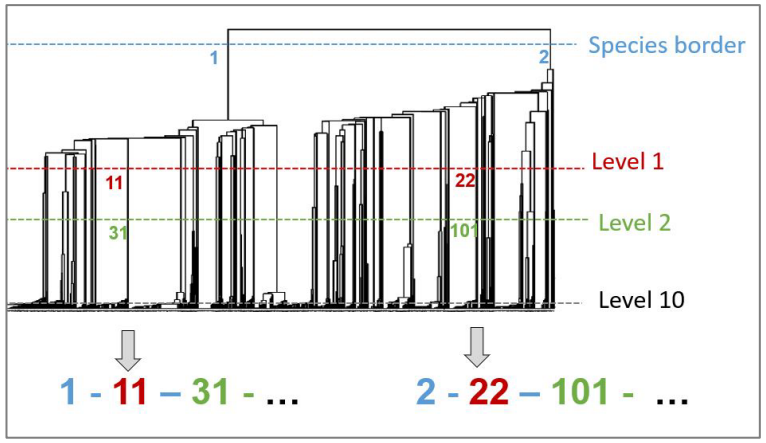
- Use of LIN codes taxonomy to be prefered over MLST

Taxonomy: requirements

Taxonomy component	Needs
Classification	• Stability • Phylogenetic compatibility • Multiple scales (epidemiology, population biology)
Nomenclature	• Automatizable • Human readable • Backwards compatible (inheritability)
Identification	• Accurate • Accessible • Fast, user friendly

Klebsiella pneumoniae identification based on LIN codes

https://bigsdb.pasteur.fr



MLST								Ribosomal MLST			scgMLST629_S				
gapA	infB	mdh	pgi	phoE	rpoB	tonB	ST	rST	species	subspecies	scgST	LINcode	Phylogroup	Sublineage	Clonal group
2	1	1	1	9	4	12	23	19197	Klebsiella pneumoniae		9677	0_0_429_0_42_0_1_0_0_0	Kp1	SL23	CG23
2	1	1	1	9	4	12	23	19197	Klebsiella pneumoniae		9678	0_0_429_0_61_0_0_0_0_0	Kp1	SL23	CG23
2	1	1	1	9	4	12	23	43589	Klebsiella pneumoniae		9679	0_0_429_0_42_1_0_0_0_0	Kp1	SL23	CG23
1	1	1	1	1	1	1	15	95524	Klebsiella pneumoniae		1	0_0_0_0_0_0_0_0_0_0	Kp1	SL15	CG15

LIN codes

Sublineage nomenclature

Sequence query

Please paste in your sequence to query against the database. Query sequences will be checked first for an
partial matches will be identified if an exact match is not found. You can query using either DNA or peptid

Please select locus/scheme: scgMLST629_S

Order results by: locus

Enter query sequence (single or multiple contigs up to whole genome in size)

```
>NC_016845.1 Klebsiella pneumoniae subsp. pneumoniae HS11286
chromosome, complete genome
GGTGGTCTGCCTCGCATAAAGCGGTATGAAAATGGATTGAAGCCCGGGCCGTGGATTCTACTCAACTTTC
GTCTTTCGAGAAAGACTCCGGGATCCTGAGTATTAAAAAGAAGATCTTTATATAGAGATCTGTTCTATTG
TGATCTCTTATTAGGATCGACTCTCTTTGTGGATAAGTCGGATCCGCGAAGTAAGATCAAAAGCTTAAGA
AGGATCACTATCTGTGAATGATCGGTGATCCTGGTCCGTATAAGCTGGGATCAGAATGAAGGGTTATGCA
CAGCTCAAAAACCATACTACGGTTATTCTTTGGATAACTACCGGTTGATCCAAGCTTTTAAGCATGGTTA
```

scgMLST629_S

Matching profile

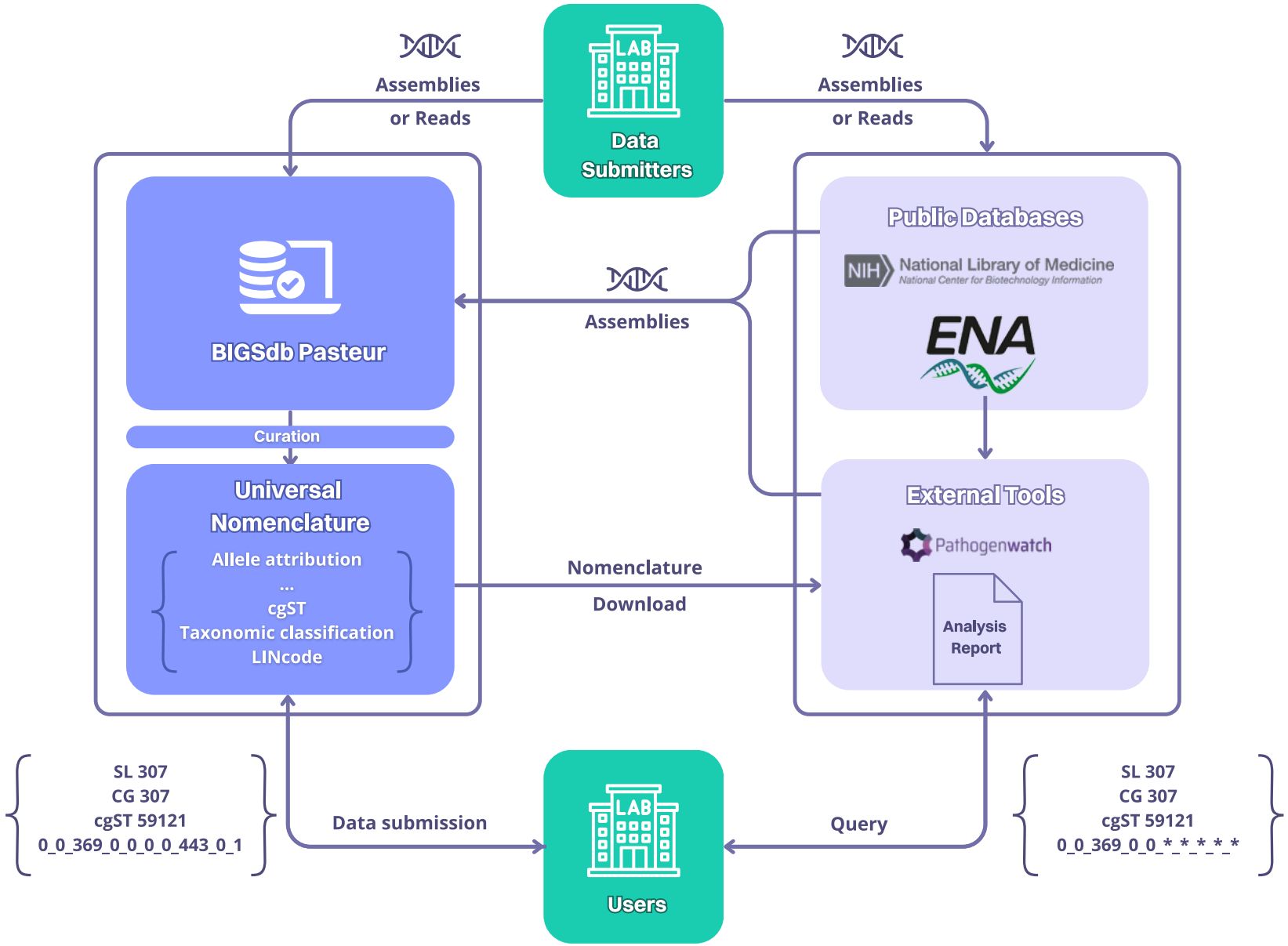
scgST: 4268

0_0_105

SL258

Genomic taxonomy ecosystem

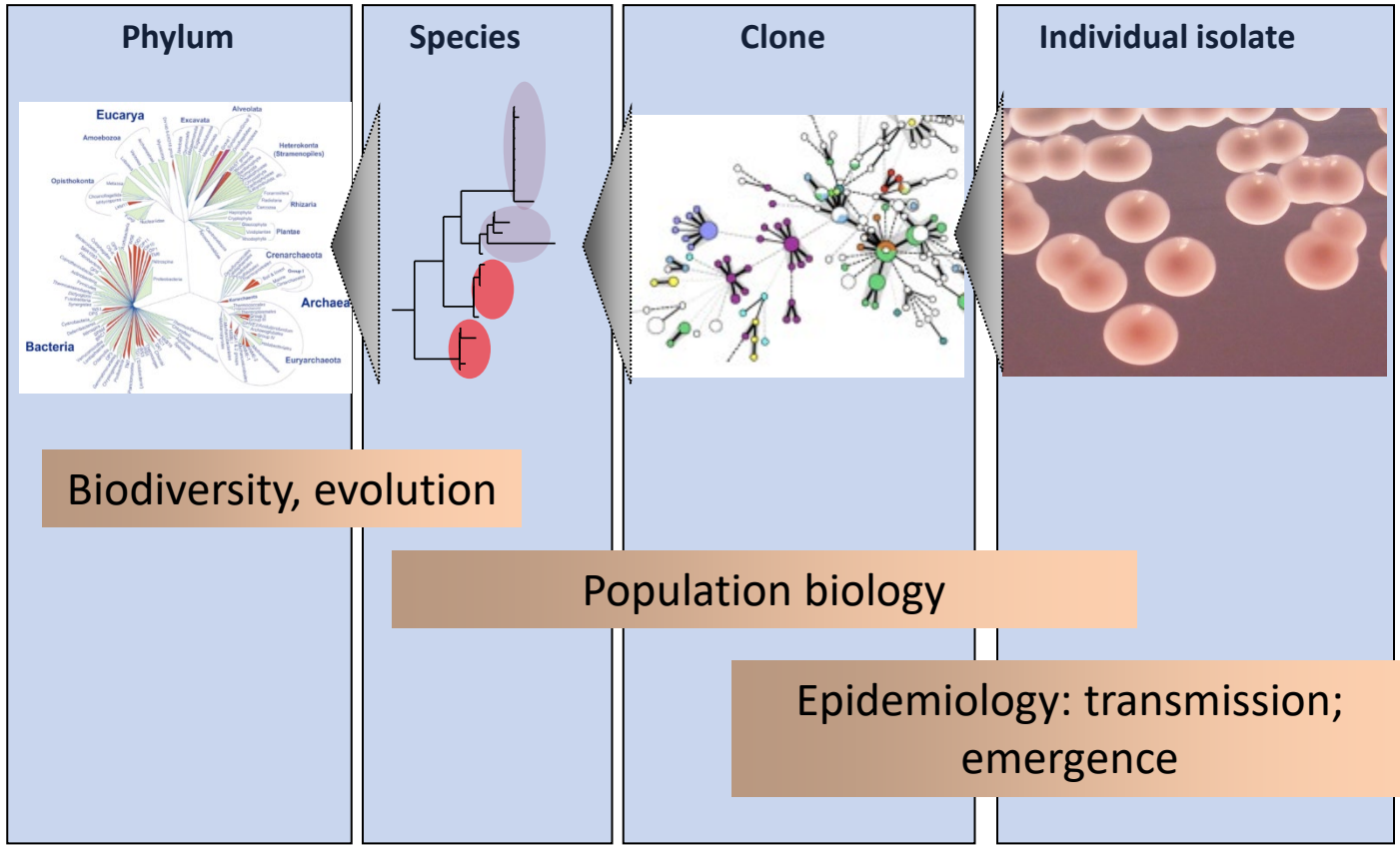
KlebNET-GSP



cgMLST LIN codes

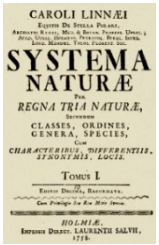
Taxonomy component	Needs
Classification	✓ Stability ✓ Phylogenetic compatibility ✓ Multiple scales (epidemiology, population biology)
Nomenclature	✓ Human readable ✓ Automatizable ✓ Backwards compatible (inheritability)
Identification	✓ Accurate ✓ Accessible ✓ Fast, easy

LIN codes strain taxonomy



- ✓ Stable classification
- ✓ Multi-purpose
- ✓ Automated
- ✓ Human readability
- ✓ Backward compatibility
- ✓ User-friendly
- ✓ Fast identification from genomes

Linnaean taxonomy ⟷ **Genomic taxonomy**


CAROLI LINNÆI
SYSTEMA NATURÆ
REGNA TRIA NATURÆ,
CLASSES, ORDINES, GENERA, SPECIES,
TOMUS I.

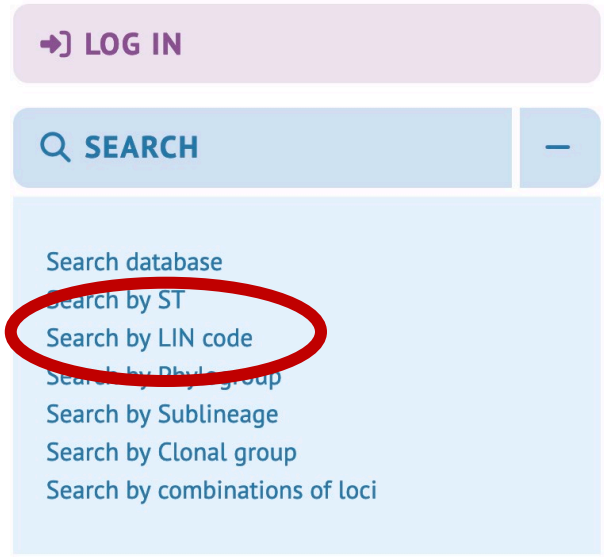
scgST	LINcode	Phylogroup	Sublineage	Clonal group
1	0_0_0_0_0_0_0_0_0_0	Kp1	SL15	CG15
9677	0_0_429_0_42_0_1_0_0_0	Kp1	SL23	CG23
9678	0_0_429_0_61_0_0_0_0_0	Kp1	SL23	CG23
9679	0_0_429_0_42_1_0_0_0_0	Kp1	SL23	CG23

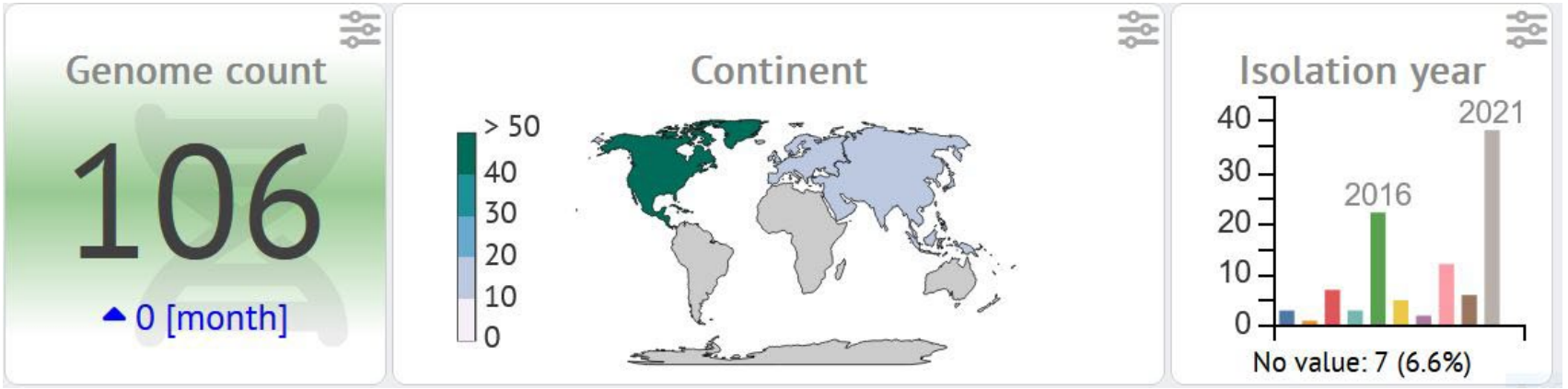
***Don't say K. pneumoniae, say K. pneumoniae* sublineage 23**

LIN codes: practicum

https://bigsdb.pasteur.fr

- How many isolates have prefix 0_0_197_0_4_1_0_0 ?
- What are their genetic diversity and geographic origins?
- How many isolates belong to clonal group 147? (CG147, of prefix 0_0_197_0)





Strains are dual-risk
(both MDR and HvKP)
Or 'convergent' phenotype

Clade B of Martin *et al*., 2021, PNAS
(Tuscany outbreak) for more details

Further reading

https://bigsdb.pasteur.fr/klebsiella/cgmlst-lincodes/ (includes a video on LIN codes)

Bacterial strain nomenclature in the genomic era: Life Identification Numbers using a gene-by-gene approach

Federica Palma, Melanie Hennart, Keith A. Jolley, Chiara Crestani, Kelly L. Wyres, Sebastien Bridel, Corin A. Yeats, Bryan Brancotte, Brice Raffestin, Sophia David, Margaret M. C. Lam, Radosław Izdebski, Virginie Passet, Carla Rodrigues, Martin Rethoret-Pasty, Martin C. J. Maiden, David M. Aanensen, Kathryn E. Holt, Alexis Criscuolo, Sylvain Brisse

doi: https://doi.org/10.1101/2024.03.11.584534

Contributors

Biodiversity & Epidemiology of Bacterial Pathogens


Virginie PASSET


Carla RODRIGUES


Sebastien BRIDEL


Mélanie HENNART

Martin RETHORET-PASTY
Audrey COMBARY


Jose Francisco DELGADO BLAS


Chiara CRESTANI


Pasteur IT & Bioinformatics HUB
Alexis CRISCUOLO


Julien GUGLIELMINI
Youssef GHORBAL, Brice RAFFESTIN
Bryan BRANCOTTE


Biological Resource Center of Institut Pasteur
Federica PALMA

Oxford University / BIGSdb tool
Keith JOLLEY
Martin MAIDEN

Oxford University / Pathogenwatch group
Sophia DAVID
Corin YEATS
David M. AANENSEN

LSHTM / HOLT group
Ebenezer FOSTER-NYARKO, Margaret LAM, Charlene RODRIGUES, Tom STANTON, Kara TSANG, Kelly WYRES, Kat HOLT

Funding