

# Bacterial strain taxonomy using LIN codes



Pr 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  
Institut Pasteur, Paris, France

September 9<sup>th</sup>, 2025

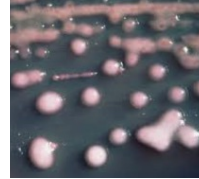


@sylvainbrisse



**Emergence and evolution  
(of bacterial strains)**

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



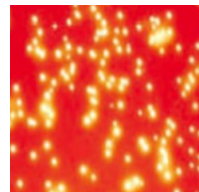
***Klebsiella pneumoniae***

Multidrug resistance



***Corynebacterium diphtheriae***

Multidrug resistance



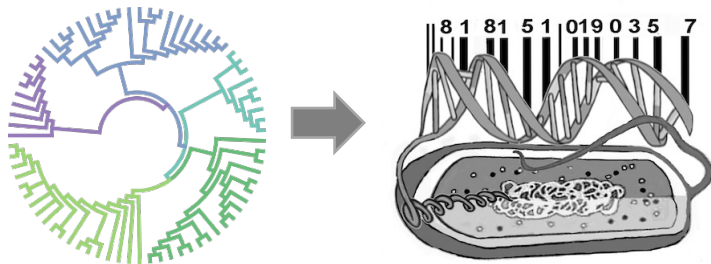
***Bordetella pertussis***

Vaccine-driven evolution

**Other related pathogens**

(of public health importance)

**Genomic taxonomies of strains**



National Reference Center



# 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



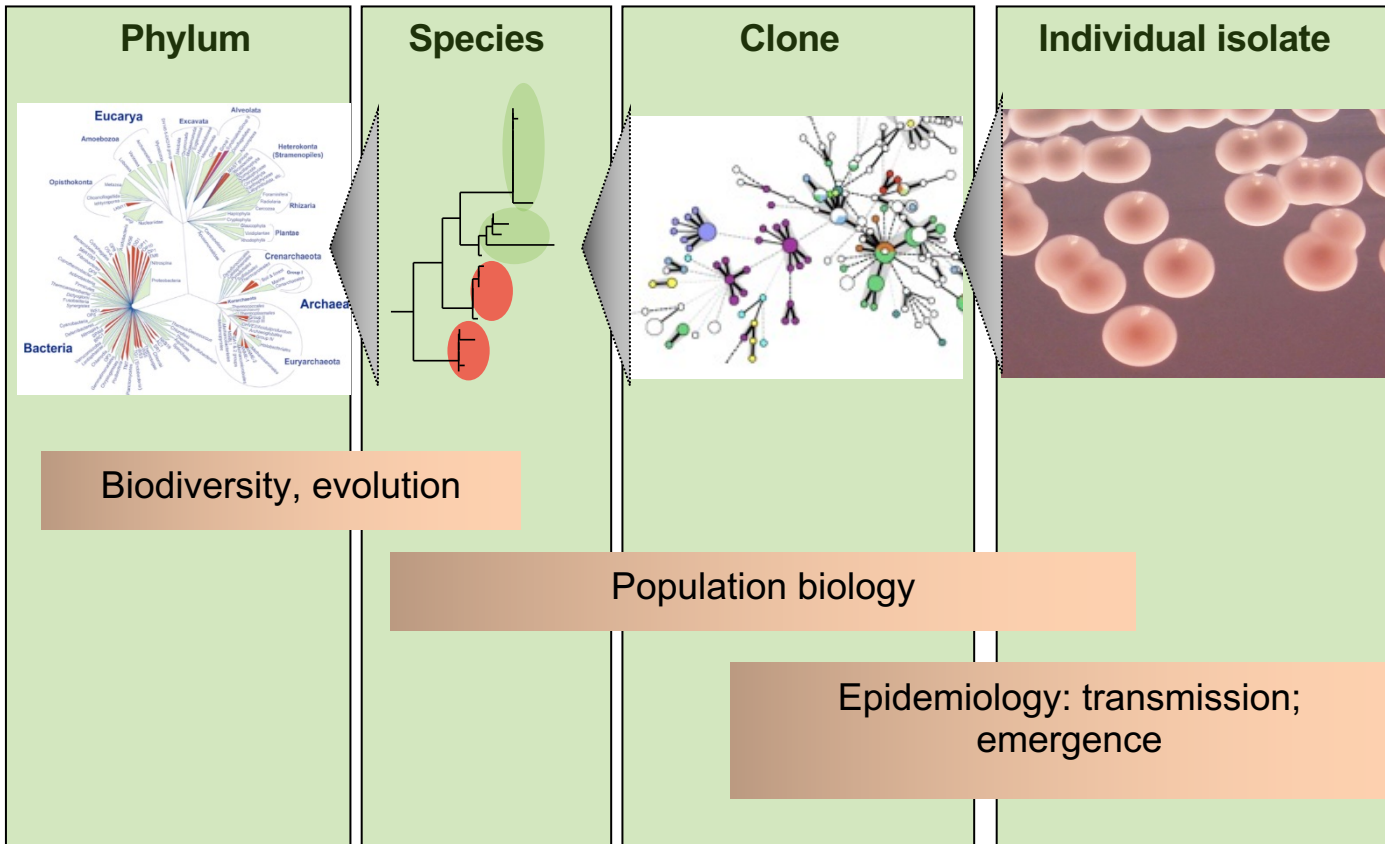
# Have you already heard about LIN codes?

① The Slido app must be installed on every computer you're presenting from



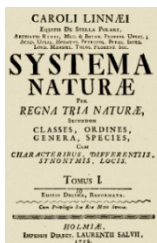
**Have you already used LIN codes  
in your surveillance or research  
activities?**

# Strain taxonomies: needs and applications

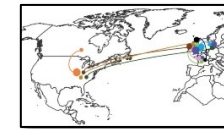


Linnaean taxonomy

Strain taxonomy



- Outbreak detection



- Global spread

- Antimicrobial resistance



- Vaccine escape

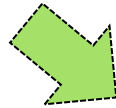


# Taxonomy

**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

Classification

- Stability
- Phylogenetic compatibility
- Multiple scales (epidemiology, population biology)

Nomenclature

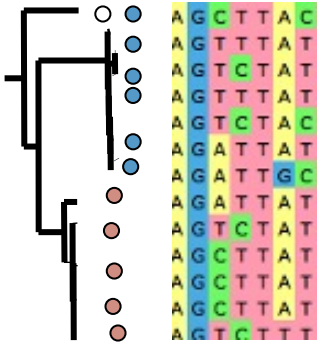
- Automatizable
- Human readable
- Backwards compatible (inheritability)

Identification

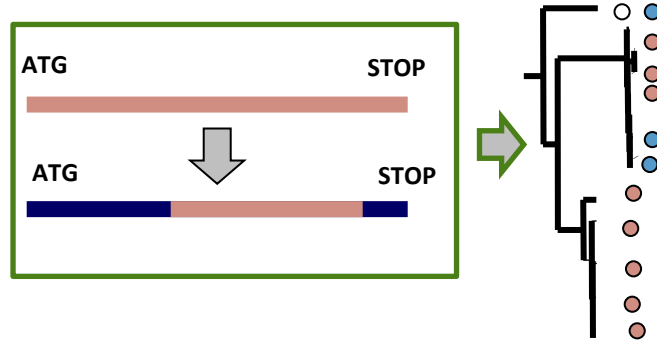
- Accurate
- Accessible
- Fast, user friendly

# Bacterial evolution: drivers & implications

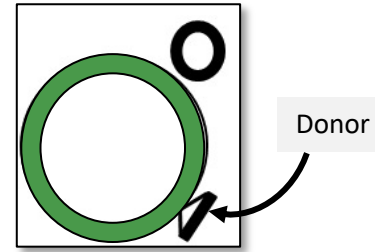
## Mutation



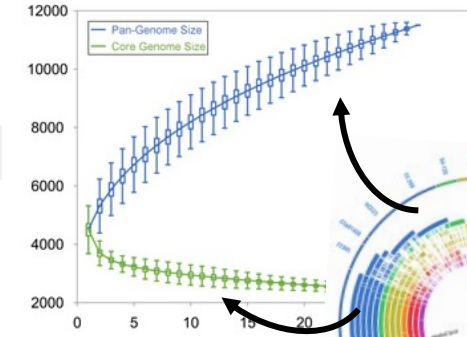
## Recombination



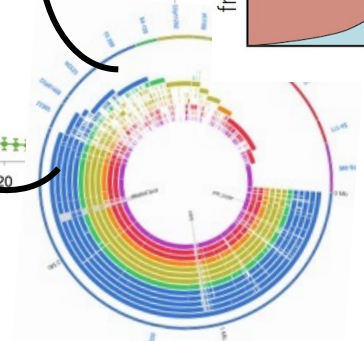
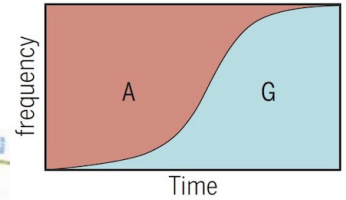
## Gene transfer



## 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

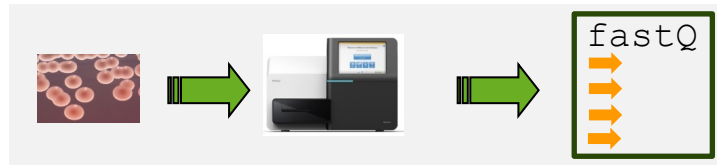
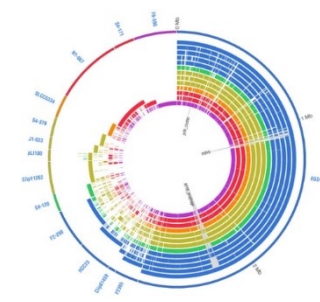
## Accessory genes:

- Subject to strong selection & transfer

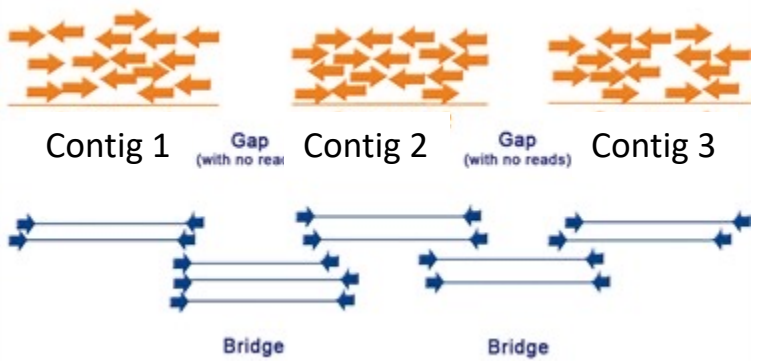
- Virulence elements
- Resistance elements
- Ecological adaptations

# How do we compare bacterial genomes?

## Genomic data analysis strategies



### De-novo assembly



Gene 1 Gene 2 Gene 3 Gene 4 ... Gene X

310 320 330 340 350

G A A A A A T C C C G T G C A T T T G C C G A T A T C G T C A A A A T C G G T C G T A C C C A

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

Gene-by-gene approaches (MLST)

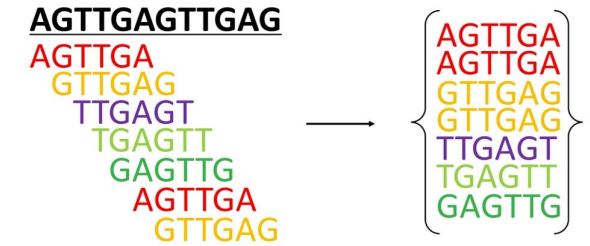
### Mapping

```

161080 161090 161100 161110 161120 161130
3ggttaccagaacatggcgggcaaacaggaacgccgggttcacgcgcataatcgttatggata
3ggttaccagaacatggcgggcaaacaggaacgccgggt
3ggttaccagaacatggcgggcaaacaggaacgccggg
3ggttaccagaacatggcgggcaaacaggaacgcc
3ggttaccagaacatggcgggcaaacaggaacgcc
3ggttaccagaacatggcgggcaaacaggaacgcc
ACCAGAACATGGCGGCAAACAGGAACGCCGGGTGCA
CCAGAACATGGCGGCAAACAGGAACGCCGGGTGCAC
CAGAACATGGCGGCAAACAGGAACGCCGGGTGCACG
AGAACATGGCGGCAAACAGGAACGCCGGGTGCACGC
AGAACATGGCGGCAAACAGGAACGCCGGGTGCACGC
GAACATGGCGGCAAACAGGAACGCCGGGTGCACGCG
GAACATGGCGGCAAACAGGAACGCCGGGTGCACGCG
AACATGGCGGCAAACAGGAACGCCGGGTGCACGCGC
ACATGGCGGCAAACAGGAACGCCGGGTGCACGCGCA
ACATGGCGGCAAACAGGAACGCCGGGTGCACGCGCA
CATGGCGGCAAACAGGAACGCCGGGTGCACGCGCAT
TGGCGGCAAACAGGAACGCCGGGTGG
GGCGGCAAACAGGAACGCCGGGTGCACGC
GGCGGCAAACAGGAACGCCGGG
GGCGGCAAACAGGAACGCCGGGTGG
GGCGGCAAACAGGAACGCCGGGTGCACGCGCATATC
GCGGCAAACAGGAACGCCGGGTGCACGCGCATATCG
GCGGCAAACAGGAACGCCGGGTGCACGCGCATATCG
CGGCAAACAGGAACGCCGGGTGCACGCGCATATCG
CGGCAAACAGGAACGCCGGGTGCACGCGCATATCGT
CGGCAAACAGGAACGCCGGGTGCACGCGCATATCGT
GGCAAACAGGAACGCCGGGTGCACGCGCATATCGTT
3GGTACCAGAACATGGCGGCAAACAGGAACGCCGGGTGCACGCGCATATCGTTATGGATA
    
```

Single Nucleotide polymorphisms (SNPs)

### K-mers



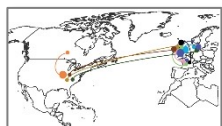
- Count & compare
- Derive genetic distance (e.g., MASH distance)

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)



International



Historical collections

MLST  
(7 genes)



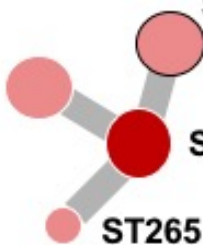
ST	gapA	infB	mdh	pgi	phoE	rpoB	tonB
14	1	6	1	1	1	1	1
15	1	1	1	1	1	1	1
265	3	1	1	1	1	1	1



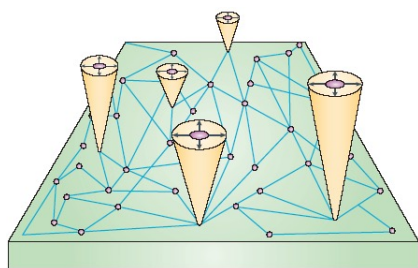
ST14 : 1 - 6 - 1 - 1 - 1 - 1 - 1

ST15 : 1 - 1 - 1 - 1 - 1 - 1 - 1

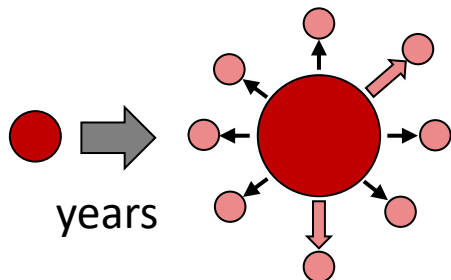
ST265 : 3 - 1 - 1 - 1 - 1 - 1 - 1



Network analyses  
(epidemiology)

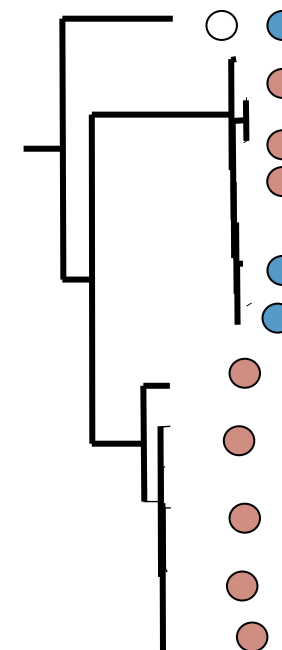


Clonal expansions



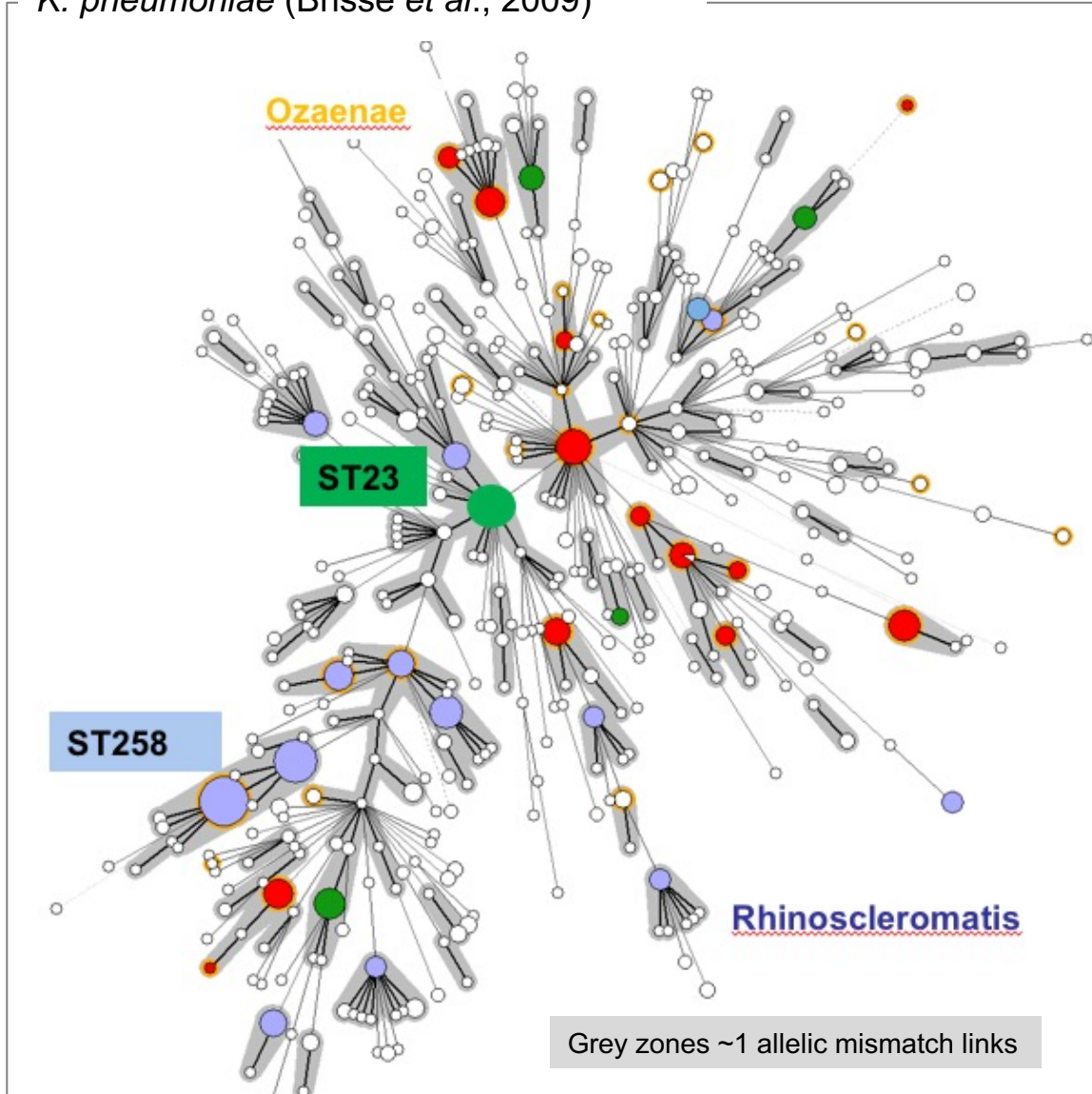
years

Phylogenetic analyses  
(evolution)



# 7-gene MLST : lack of resolution

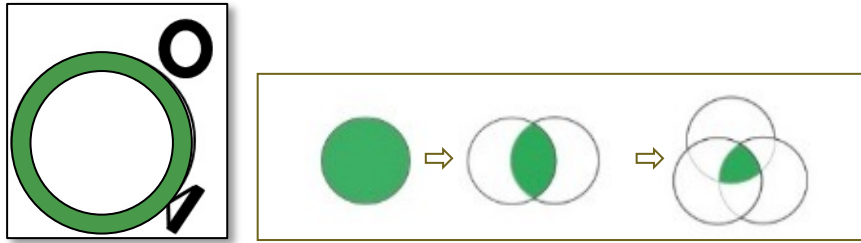
*K. pneumoniae* (Brisse *et al.*, 2009)



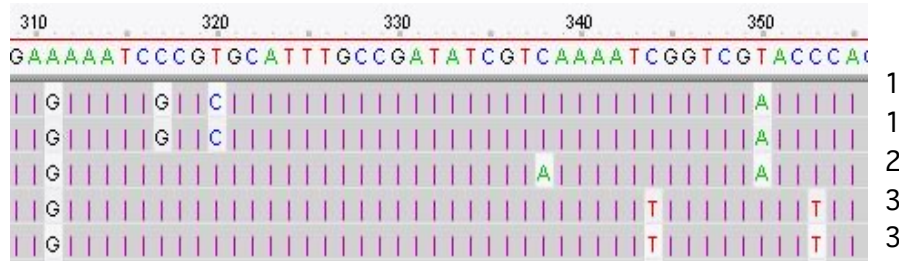
- 7 genes ~ 3500 nucleotides
  - ~ 1/1000th of genome length
  - Most STs are globally distributed
  - STs can be > 100 years old
  - ST do not inform on recent transmission
- ... but MLST is a common language



# Core genome MLST



**Step 1.** Define core genome



**Step 2.** Define variation at core genes

1	15	12	37	3	16	22	11	1	15	12	37
1	5	4	24	4	4	1	1	1	5	4	24
27	28	9	31	1	25	10	15	27	28	9	31



1000-2000 gene loci

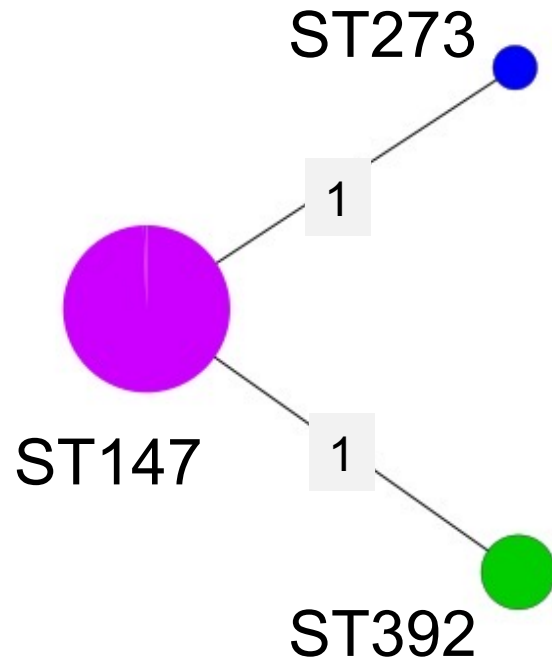
**Step 3.** Define allelic profiles of genomes

- Discrimination
  - Classification precision
- are much improved compared to classical (7-gene) MLST

# *Klebsiella pneumoniae* clonal group 147

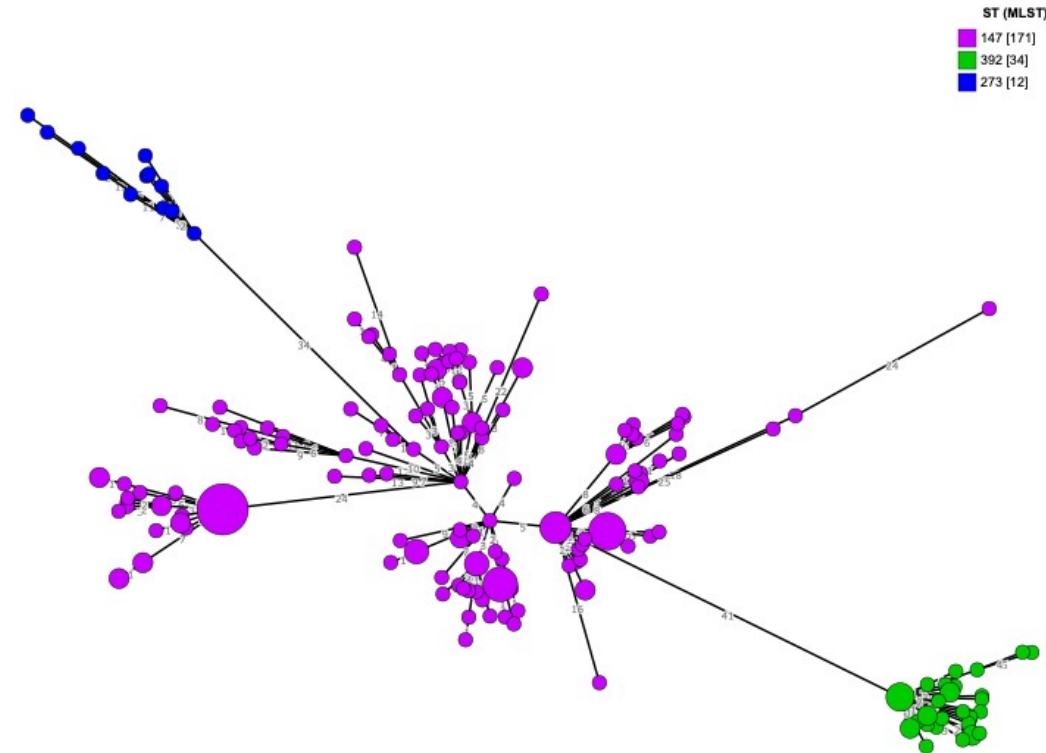
## MStree based on MLST (7 genes)

(Diancourt et al., J Clin Microbiol 2005)

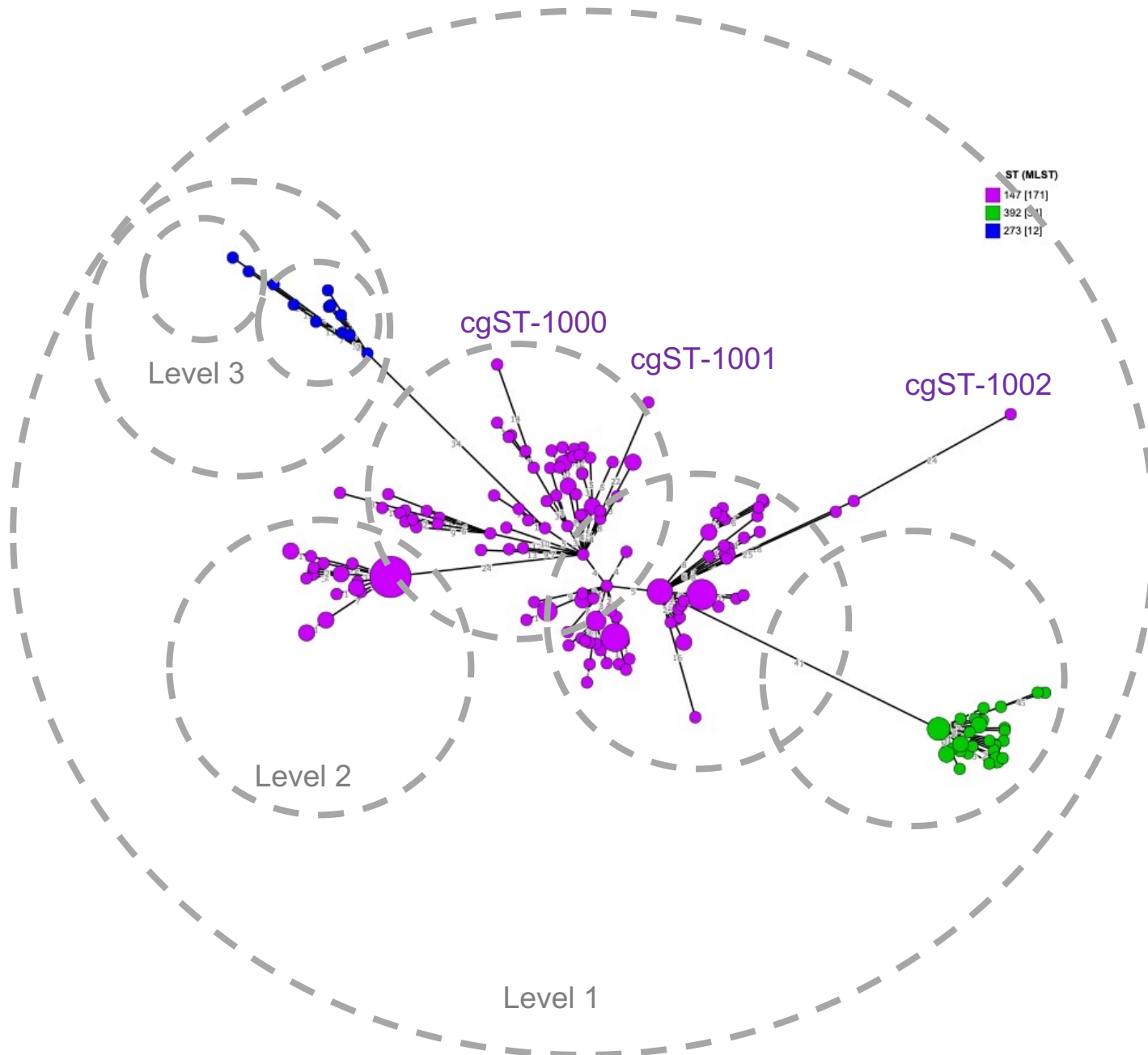


## MStree based on cgMLST (629 genes)

(Hennart et al. Mol Biol Evol 2022)



# How can we classify bacteria using cgMLST profiles?

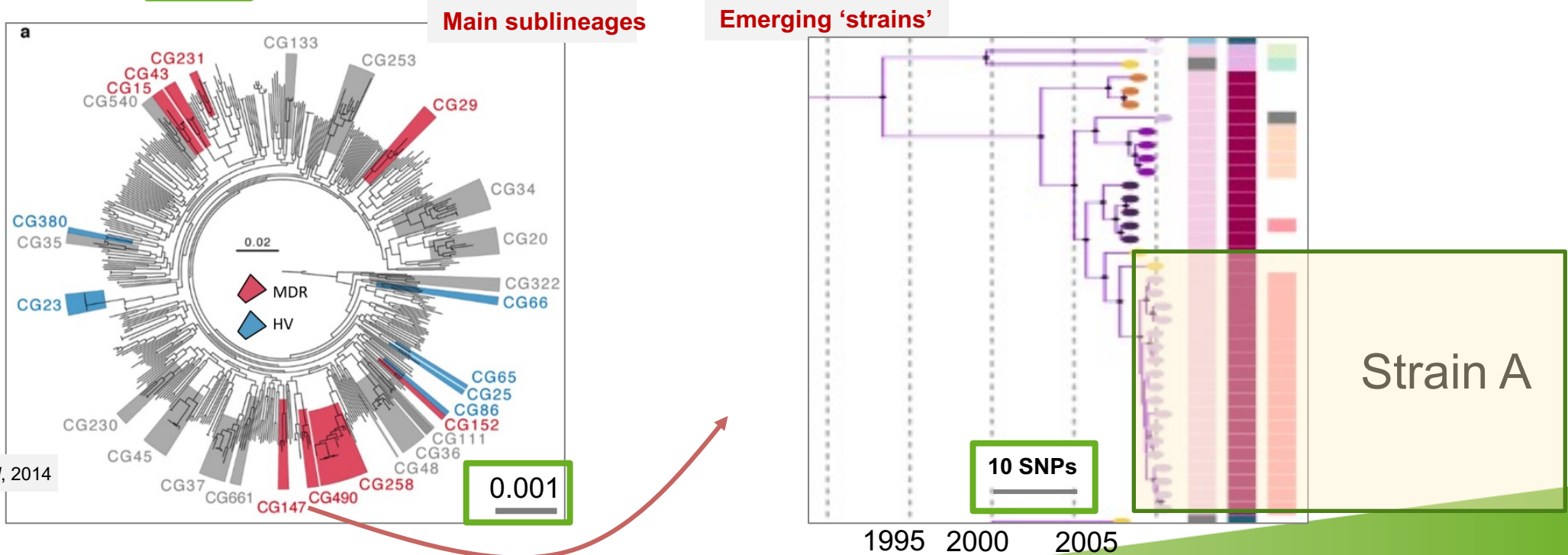
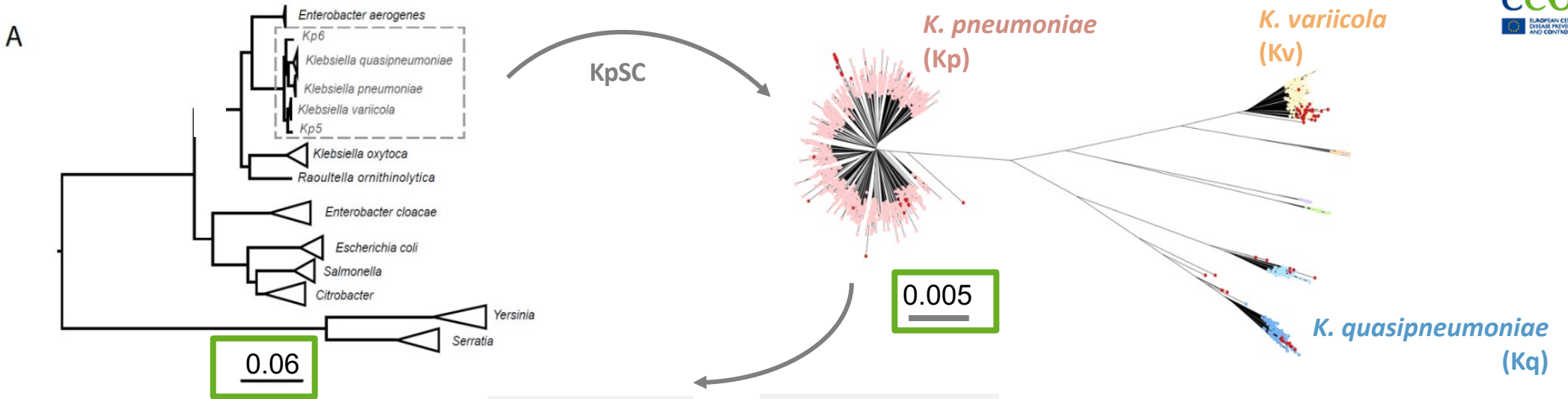


- **Problem:** cgSTs are too discriminatory for meaningful epidemiological associations
- **Solution:** group cgSTs based on their profiles' similarity levels

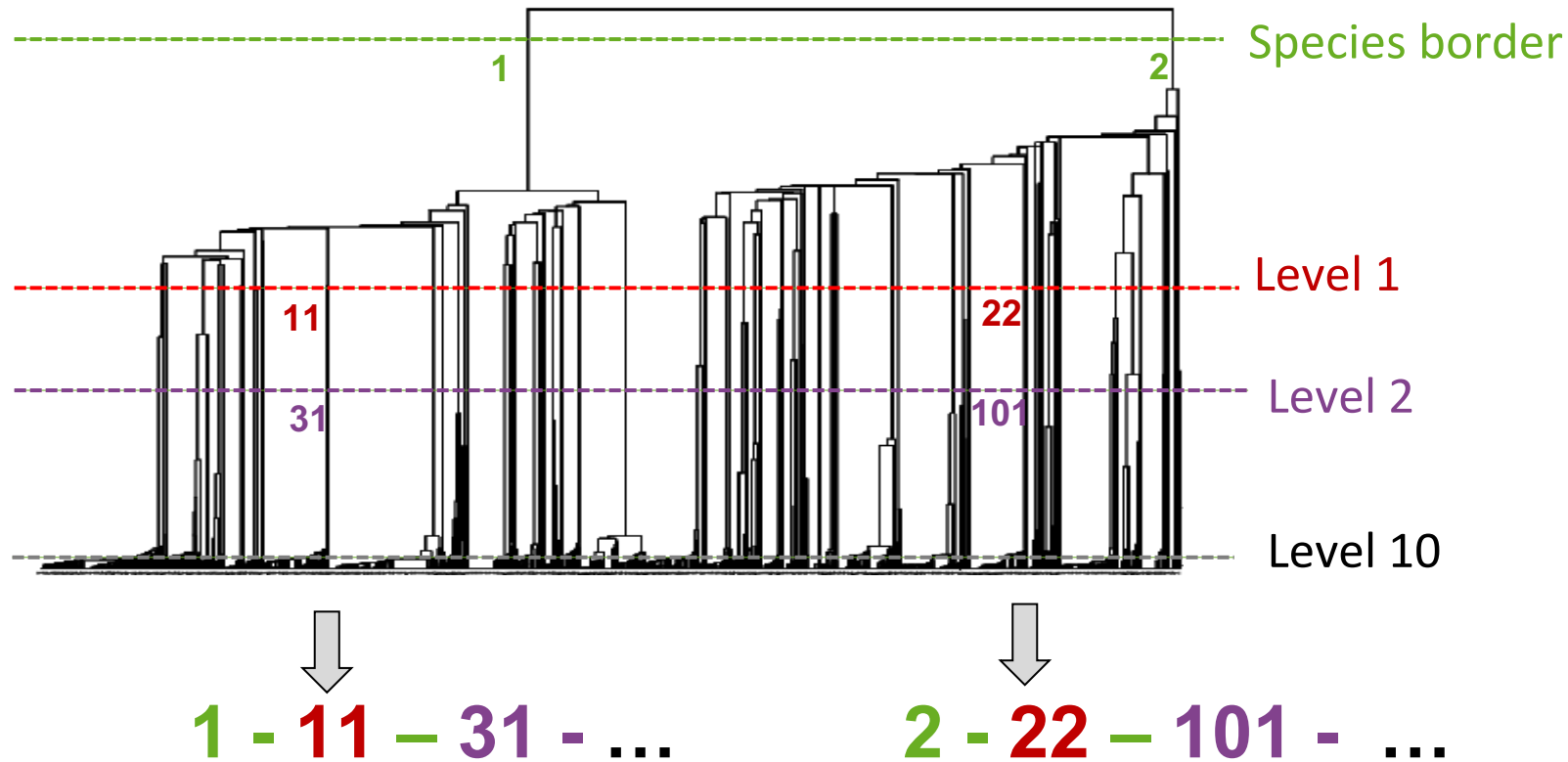
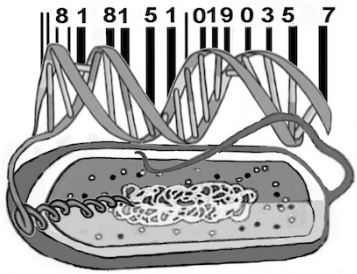


Need to choose relevant phylogenetic depths

# The multiple levels of phylogenetic structure

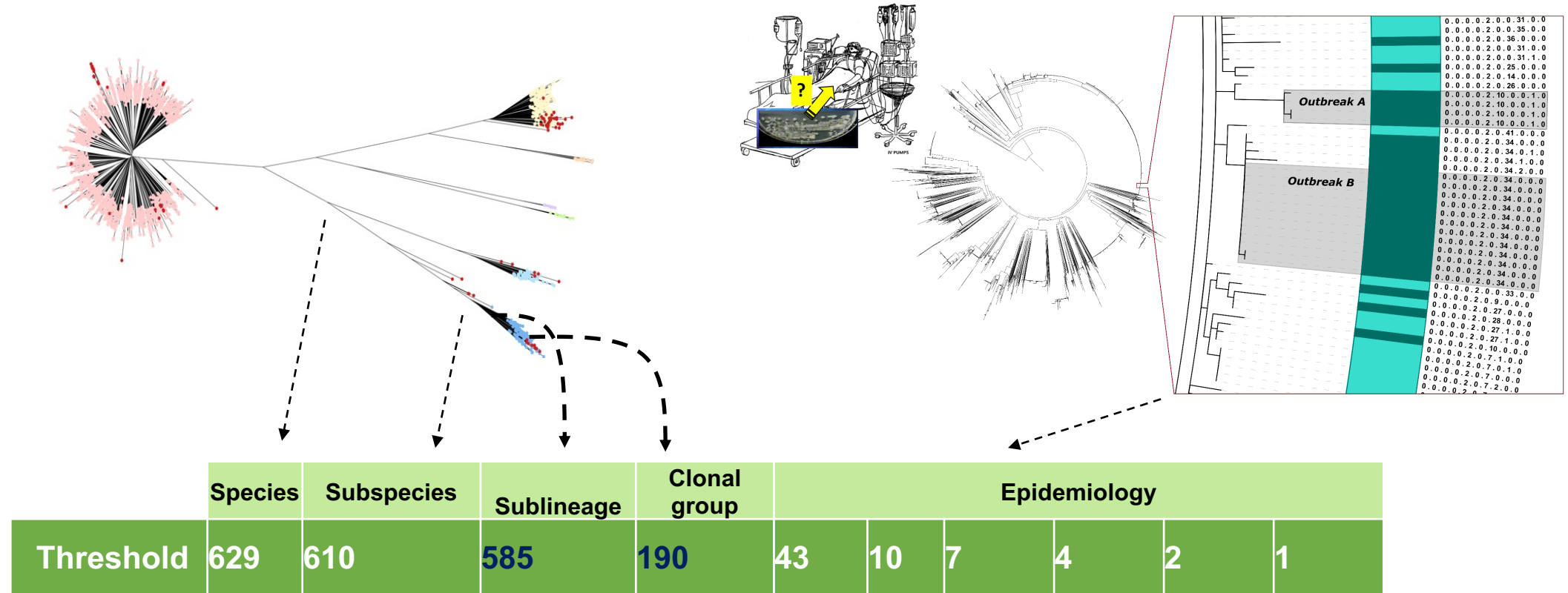


# Multilevel classification



➔ 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



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					
<b>Threshold</b>	629	610	<b>585</b>	<b>190</b>	43	10	7	4	2	1

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



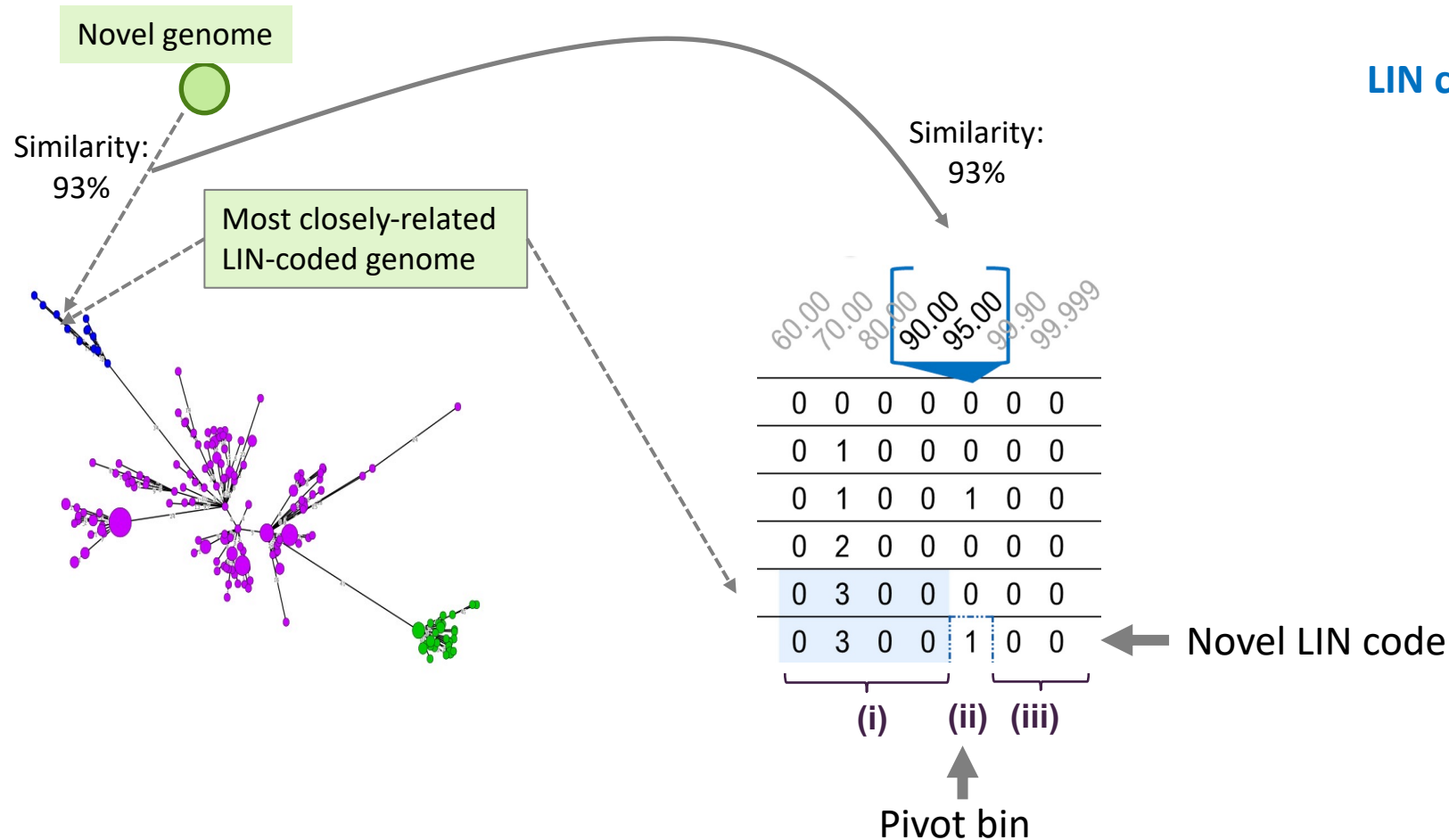
Palma, Hennart *et al.*, bioRxiv, 2024

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:

- Prefix up to pivot bin
- +1 in pivot bin
- 0 in downstream bins

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



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:										
<b>Closest genome (similarity %)</b>		0	3.02	6.99	69.79	93.16	98.41	98.88	99.36	99.68	99.84	100
<b>Genome A</b>	Initialization	0	0	0	0	0	0	0	0	0	0	
<b>Genome B</b>	<b>A</b> (3.50%)	0	1	0	0	0	0	0	0	0	0	
<b>Genome C</b>	<b>B</b> (99.0%)	0	1	0	0	0	0	1	0	0	0	
<b>Genome D</b>	<b>B</b> (7.00%)	0	1	1	0	0	0	0	0	0	0	
...	...	...										
<b>Genome E</b>	<b>D</b> (5.00%)	0	2	0	0	0	0	0	0	0	0	
<b>Genome F</b>	<b>E</b> (98.90%)	0	2	0	0	0	0	1	0	0	0	

\* 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

Coding steps (genome E):



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

		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
<b>Genome A</b>	Initialization		0	0	0	0	0	0	0	0	0	0	
<b>Genome B</b>	<b>A</b> (3.50%)		0	1	0	0	0	0	0	0	0	0	
<b>Genome C</b>	<b>B</b> (99.0%)		0	1	0	0	0	0	1	0	0	0	
<b>Genome D</b>	<b>B</b> (7.00%)		0	1	1	0	0	0	0	0	0	0	
...	...												
<b>Genome E</b>	<b>D</b> (5.00%)		0	2	0	0	0	0	0	0	0	0	
<b>Genome F</b>	<b>E</b> (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)



# LIN codes levels are nested from right to left

LIN prefix, size 2	LIN prefix, size 3	LIN prefix, size 4
0_0	0_0_0	0_0_105_6
1_0	0_0_429	0_0_105_0
1_1	0_0_105	0_0_105_2
2_0	0_0_158	0_0_105_1
2_1	0_0_197	1
3_0	0_0_369	0_0_105_1
4_0	0_0_750	0_0_105_2
		9
		0_0_105_7

LIN prefix
0_0_1_0
0_0_2_0
0_0_105_0



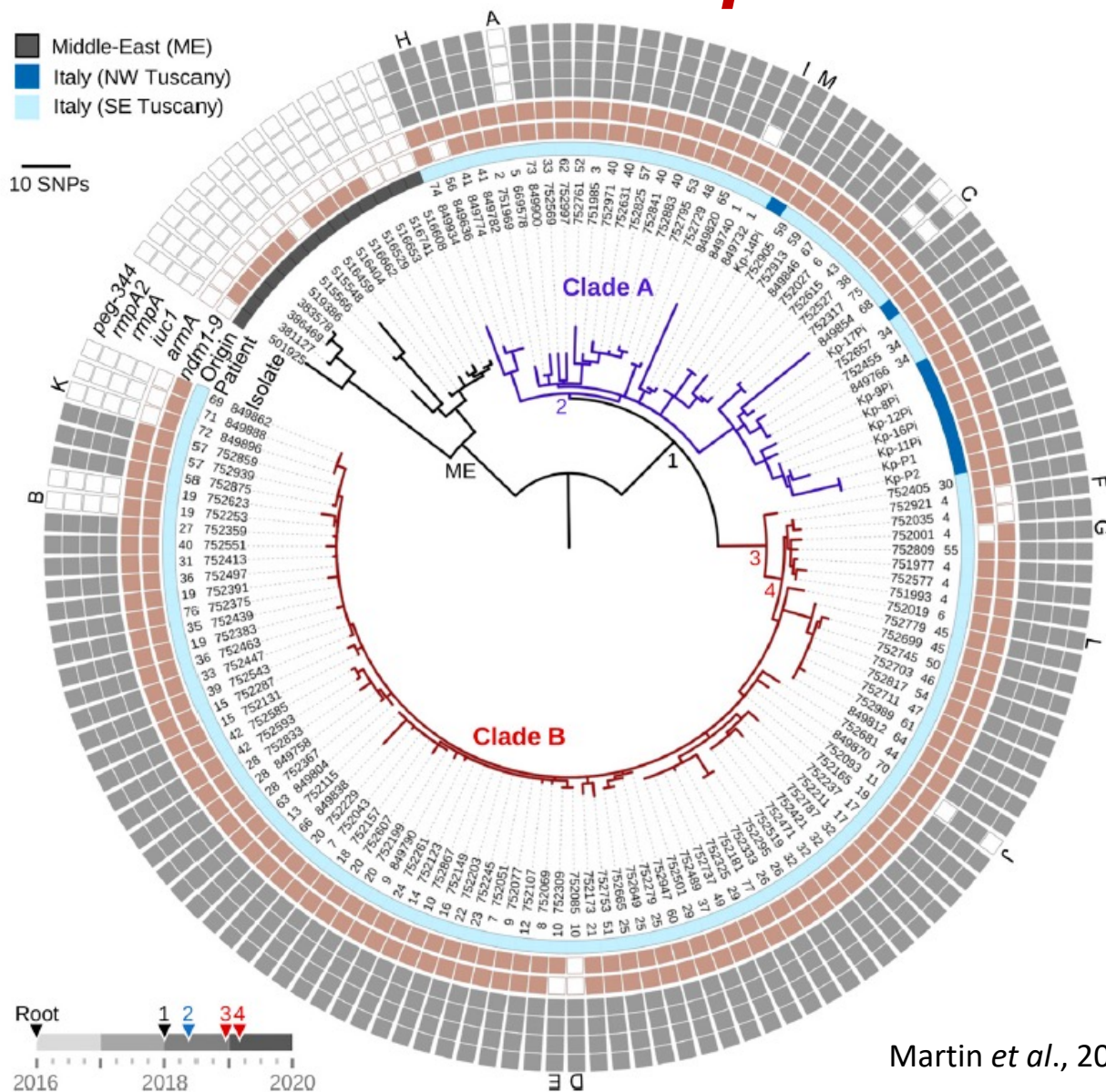
0 has a different meaning  
in these three lines

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

LIN codes are predominantly composed of small integers

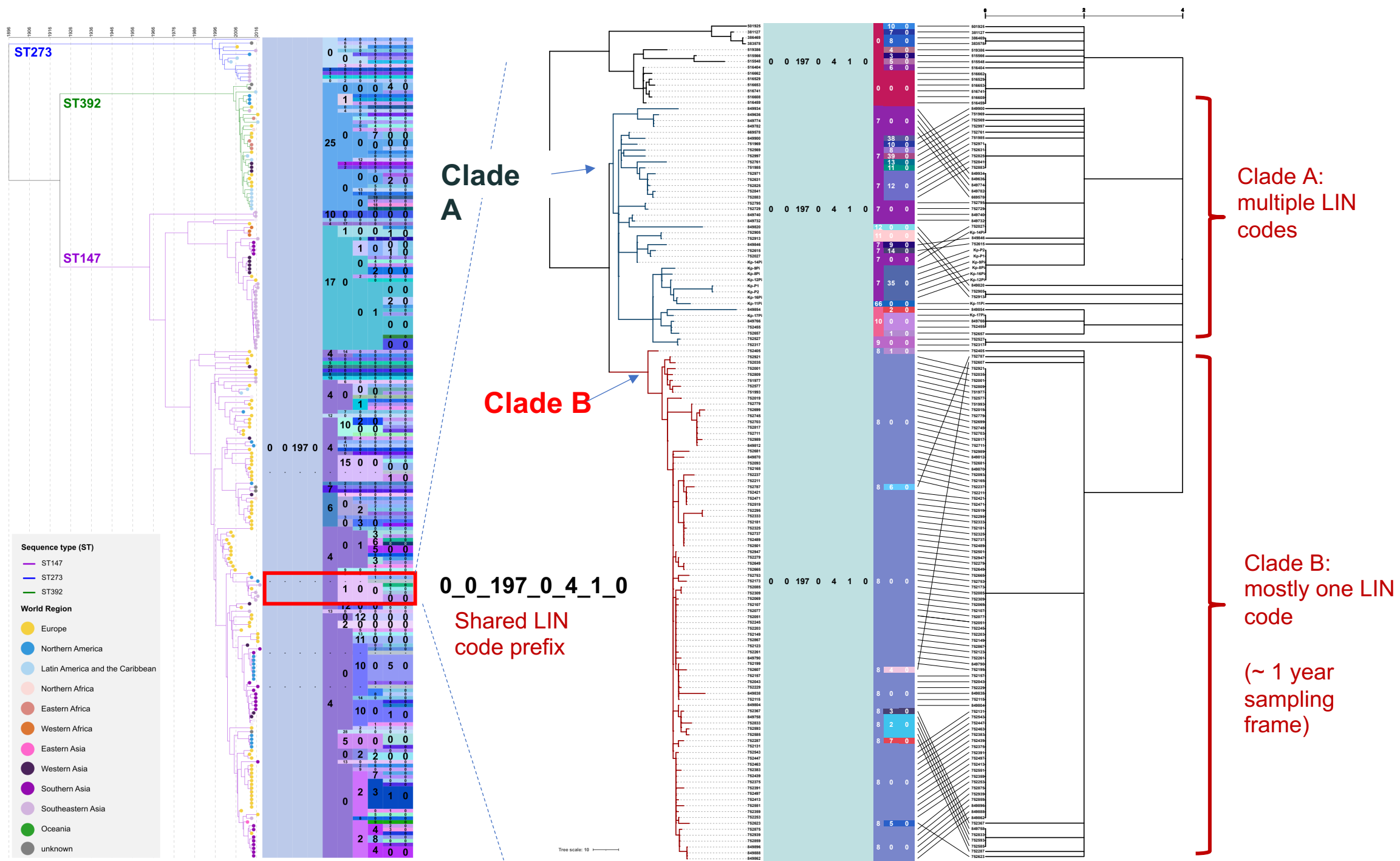


# Anatomy of an extensively drug-resistant *Klebsiella pneumoniae* outbreak



- 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



# LIN codes of Italian SL147 outbreak

# Taxonomy: requirements



## Taxonomy component

Classification

Nomenclature

Identification

## Needs

- Stability
- Phylogenetic compatibility
- Multiple scales (epidemiology, population biology)
  
- Automatizable
- Human readable
- Backwards compatible (inheritability)
  
- 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	LIN prefix	Main ST	Nickname	LIN prefix	Main ST	Nickname
0_0	Kp1	0_0_0	15	SL15	0_0_105_6	258	CG258
1_0	Kp3	0_0_429	23	SL23	0_0_105_0	340	CG340
1_1	Kp5	0_0_105	258	SL258	0_0_105_2	11	CG11
2_0	Kp2	0_0_158	45	SL45	0_0_105_11	11	CG3666
2_1	Kp4	0_0_197	147	SL147	0_0_105_1	437	GC10268
3_0	Kp6	0_0_369	307	SL307	0_0_105_29	11	CG12811
4_0	Kp7	0_0_750	6589	SL10691	0_0_105_7	895	CG895



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



# Taxonomy: requirements



## Taxonomy component

Classification

Nomenclature

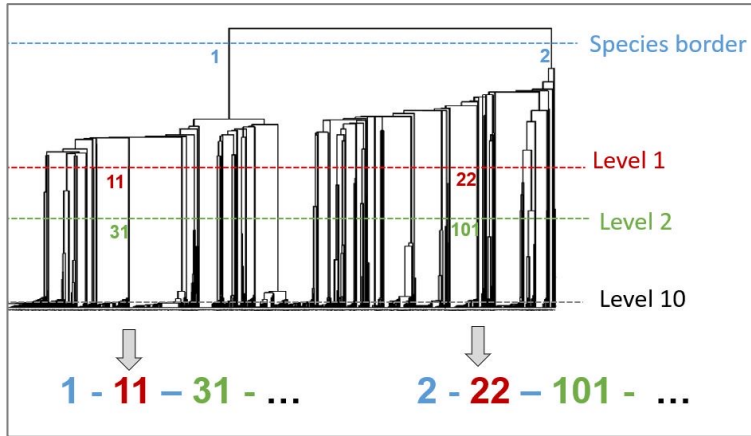
Identification

## Needs

- Stability
- Phylogenetic compatibility
- Multiple scales (epidemiology, population biology)
  
- Automatizable
- Human readable
- Backwards compatible (inheritability)
  
- 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

## 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

Order results by

scgMLST629\_S

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
GTCITTCGAGAAAGACTCCGGGATCCTGAGTATTAAGAAAGATCTTTATATAGAGATCTGTTCTATTG
TGATCTCTATTAGGATCGACTCTTTGTGGATAAGTCGGATCCGCGAAGTAAGATCAAAAGCTTAAGA
AGGATCACTATCTGTGAATGATCGGTGATCCTGGTCCGTATAAGCTGGGATCAGAATGAAGGGTTATGCA
CAGCTCAAAAACCATACGTTATTCTTTGGATAACTACCGTTGATCCAAGCTTTAAGCATGGTTA
```



scgMLST629\_S



Matching profile

scgST: 4268

LIN codes

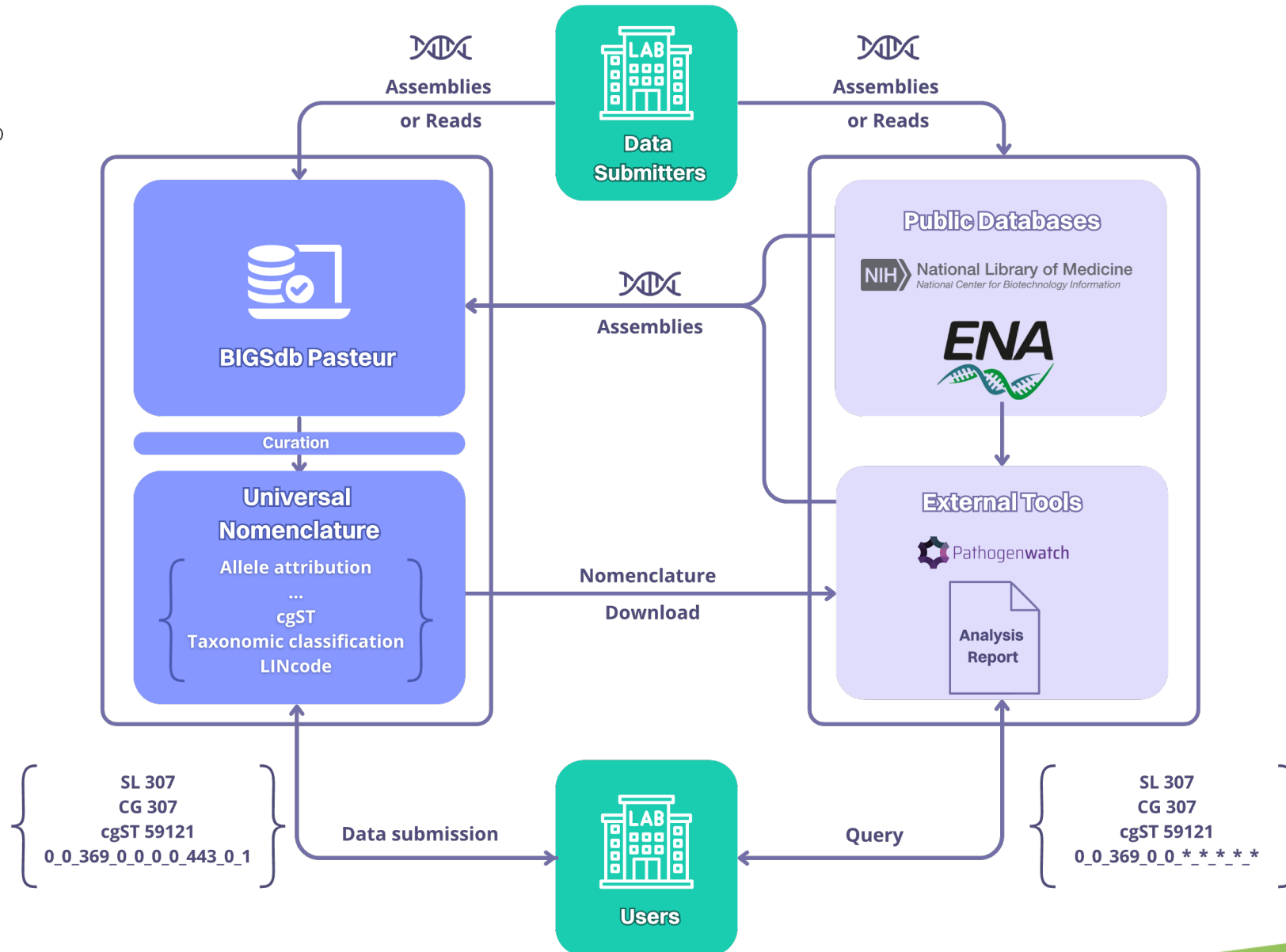
0\_0\_105

Sublineage nomenclature

SL258



# Genomic taxonomy ecosystem



# cgMLST LIN codes

## Taxonomy component

Classification

Nomenclature

Identification

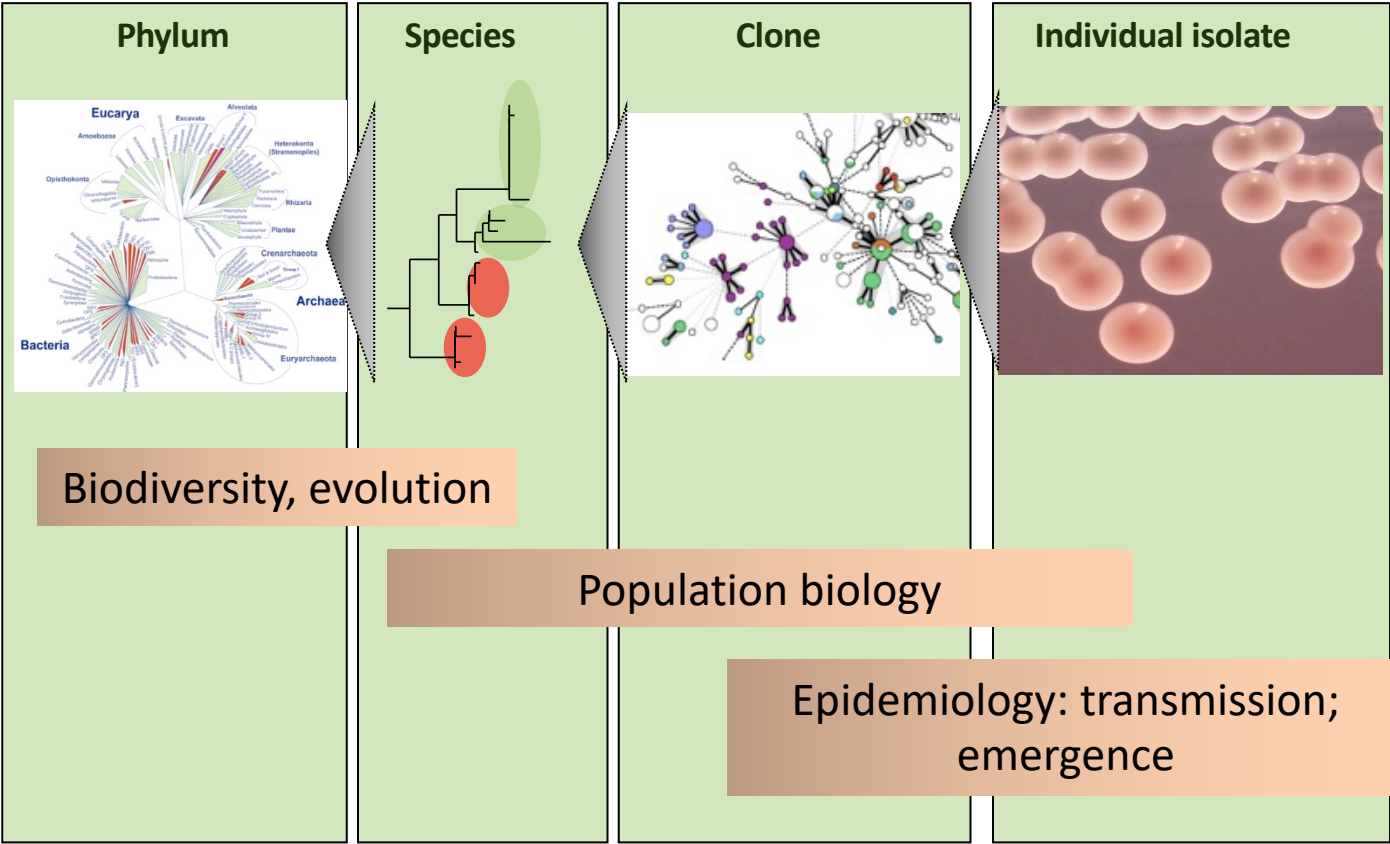
## Needs

- ✓ Stability
- ✓ Phylogenetic compatibility
- ✓ Multiple scales (epidemiology, population biology)

- ✓ Human readable
- ✓ Automatizable
- ✓ Backwards compatible (inheritability)

- ✓ Accurate
- ✓ Accessible
- ✓ Fast, easy

# LIN codes strain taxonomy

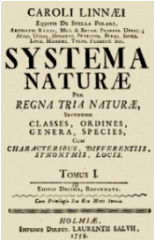


- ✓ Stable classification
- ✓ Multi-purpose
- ✓ Automated
- ✓ Human readability
- ✓ Backward compatibility
- ✓ User-friendly
- ✓ Fast identification from genomes



Linnaean taxonomy

Genomic taxonomy



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://bigbdb.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)

→] LOG IN

Q SEARCH

- Search database
- Search by ST
- Search by LIN code
- Search by Phyl. group
- Search by Sublineage
- Search by Clonal group
- Search by combinations of loci



## Allele designations/scheme fields

LINcode (scgMLST629 S)

starts with

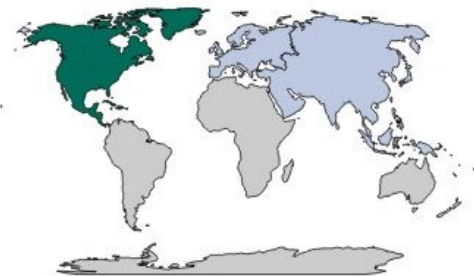
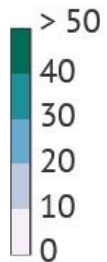
0\_0\_197\_0\_4\_1\_0\_0



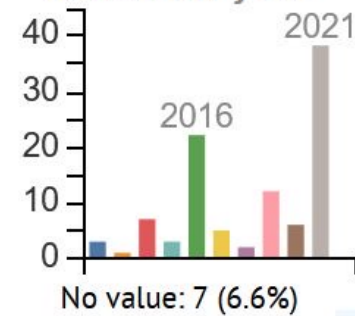
Genome count

106

▲ 0 [month]



Isolation year



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

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

# Contributors

## Biodiversity & Epidemiology of Bacterial Pathogens



Virginie PASSET



Carla RODRIGUES



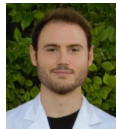
Sebastien BRIDEL



Mélanie HENNART



Chiara CRESTANI



Jose Francisco  
DELGADO  
BLAS



**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



wellcome trust

## 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



# Further reading

<https://bigsdatabase.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>



# Acknowledgements

The creation of this training material was commissioned by ECDC to Institut Pasteur with the direct involvement of Sylvain BRISSE