Bacterial strain taxonomy and 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

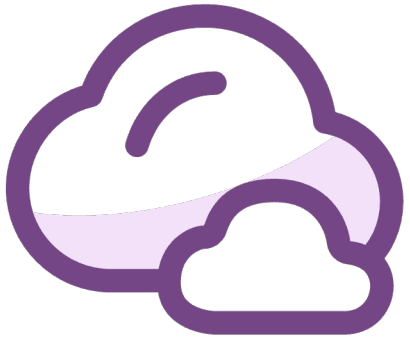
October 22nd, 2025

@sylvainbrisse

Intended Learning Objectives

Specific objectives of this session:

- Importance of strain nomenclature in public health microbiology
- Traditional *vs*. genomic taxonomy
- From species to strains: levels of resolution
- Overview of comparative genomics approaches: MLST, SNP, k-mer
- LIN codes principles
- LIN codes properties & usage



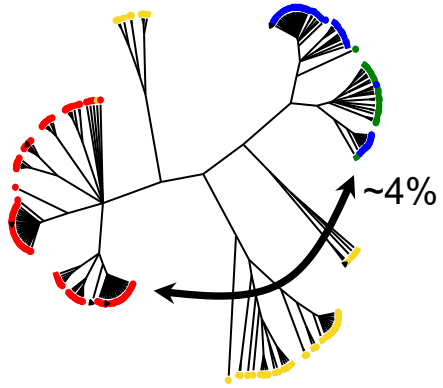
Why is strain taxonomy important? (in only two words)

Outline

- **Part 1**: Introduction to bacterial strain nomenclatures
- **Part 2**: LIN codes

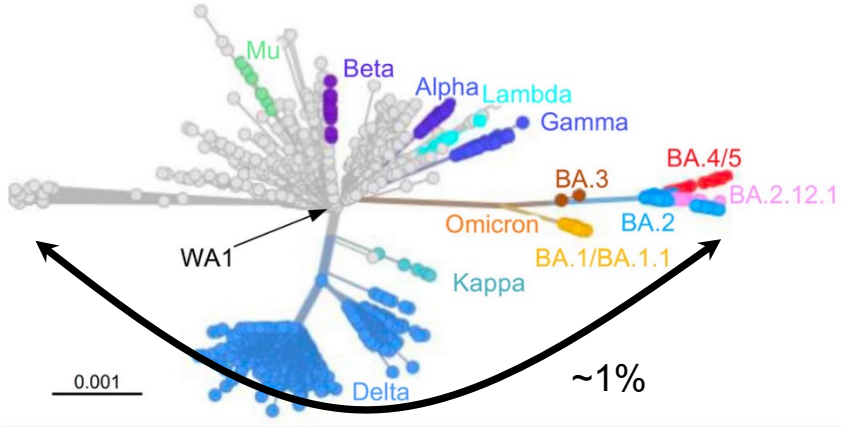
High amounts of genetic diversity within bacterial species

Species borders:
~5% DNA sequence divergence



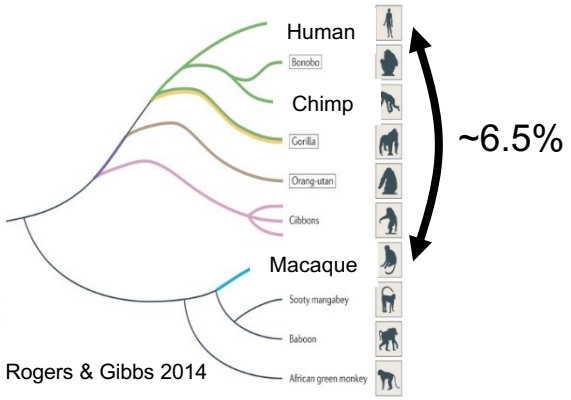
Genome: 5 MB

Genetic Distance of SARS-CoV-2 Variants



Genome: 0.03 MB

Humans - chimpanzees ~1.2%

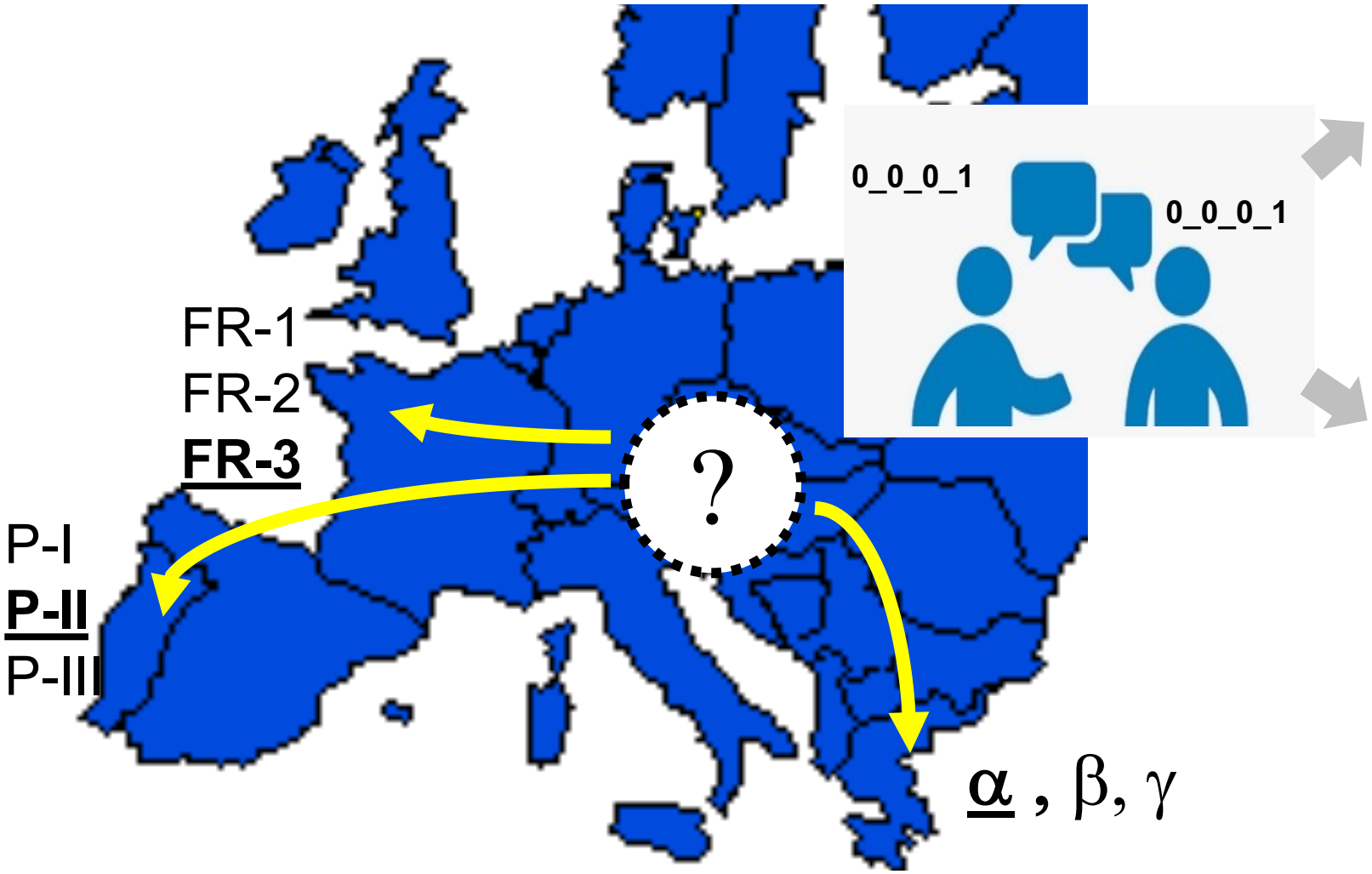


Genome: 3000 MB

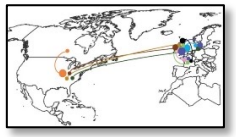
- Thousands of « strains » within bacterial species
- Bacterial strains differ by important characteristics (accessory genome)
- Bacterial populations differ across geography and time

Impacts: clinical management, diagnostics, therapies, vaccines, …

Need for a common language



- Outbreak detection



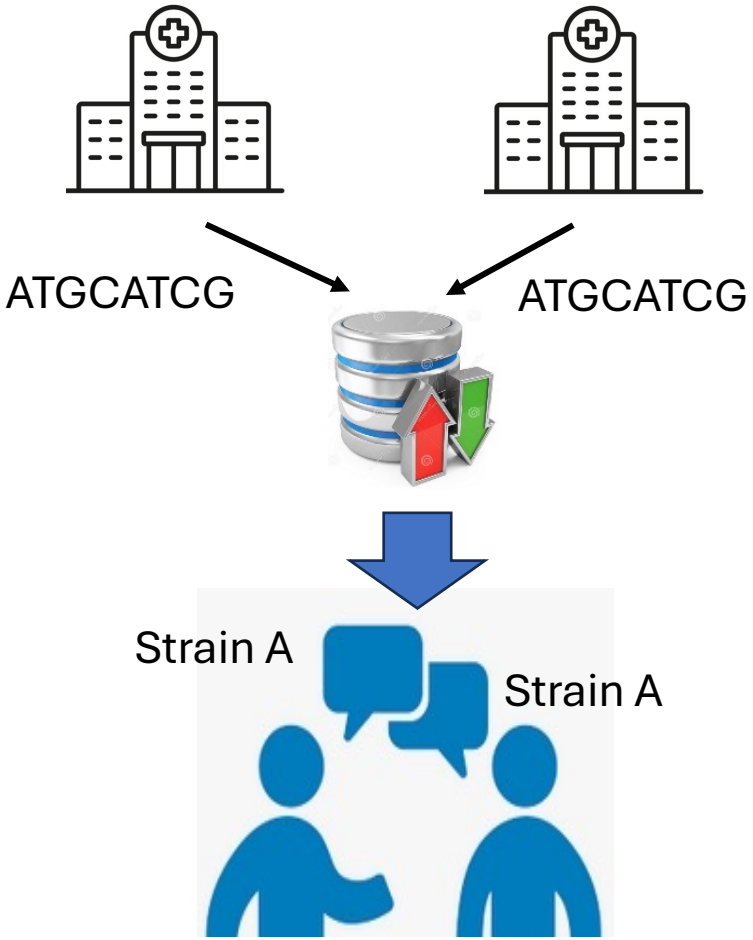
- Surveillance

- Antimicrobial-resistant sublineages

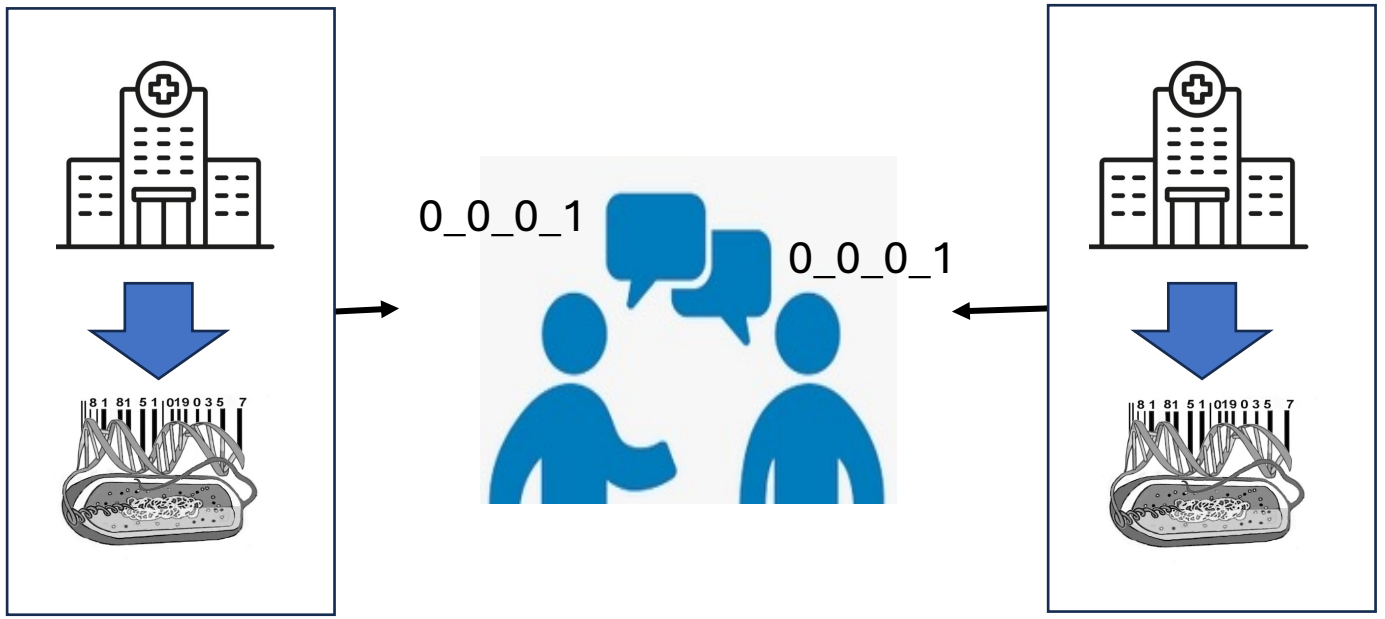
- Vaccine-escape sublineages

Two models of multi-actor surveillance

Sequence data sharing

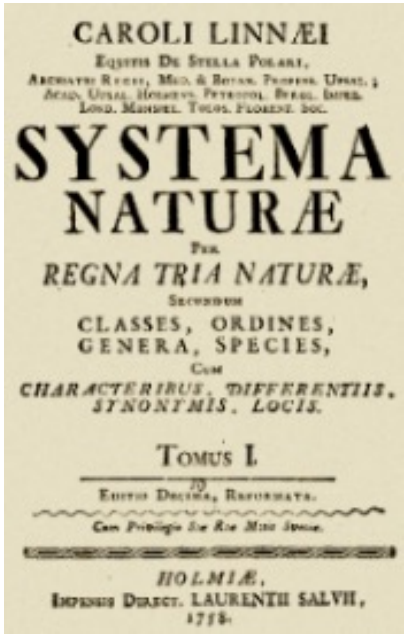


Public nomenclature



- Confidentiality of sequence data
- Portability of strain names *versus* sequence data
- Faster information exchange

Linnaean *versus* strain taxonomies


CAROLI LINNÆI

SYSTEMA NATURÆ

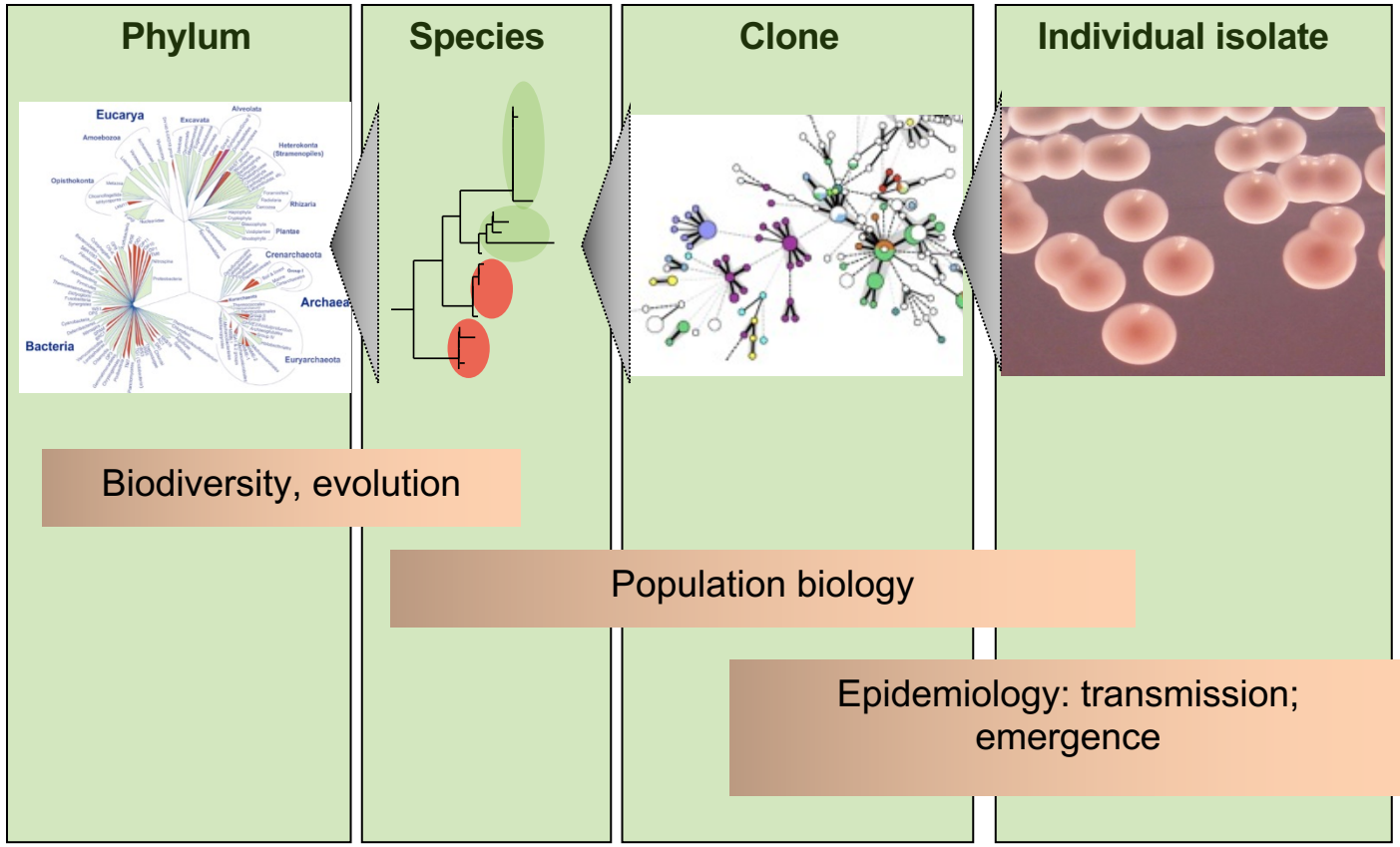
REGNA TRIA NATURÆ,

CLASSES, ORDINES, GENERA, SPECIES,

CUM CHARACTERIBUS, DIFFERENTIIS, SYNONYMIS, LOCIS.

TOMUS I.

HOLMIÆ,



Bacterial strain taxonomy: historical perspective

- Phage types, e.g. DT104
- Serotypes, e.g. Typhimurium
- PFGE types, other fingerprint-types
- MLST types (ST), clonal complex

} Pre-genomic era, low resolution

➡ Genomic sequencing-based classifications

Taxonomy

Classification
Delineating taxa

Nomenclature
Assigning names

Description
Properties of taxa

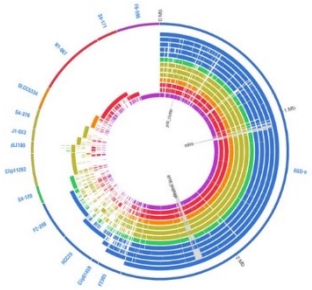
Identification
Determining to which taxon an individual belongs to
(Microbiological diagnostic)

'Biological esperanto'
'Lingua franca'

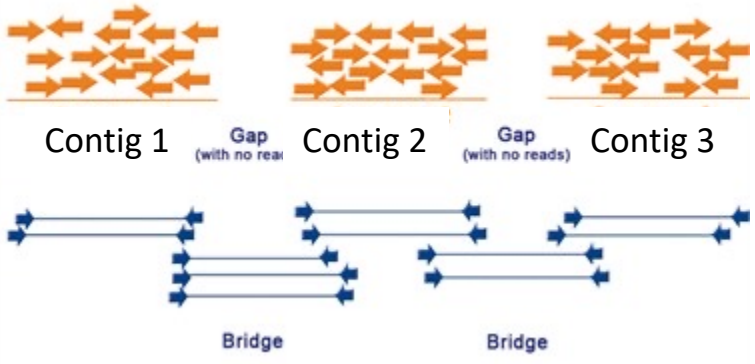
Taxonomy: requirements

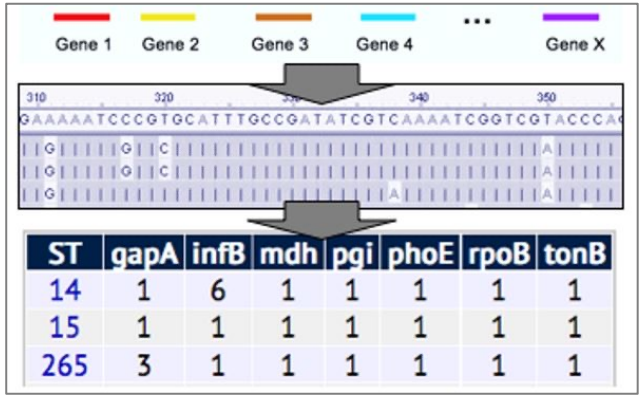
Taxonomy component	Needs
Classification	• Stability • Phylogenetic compatibility • Multiple scales (epidemiology, population biology)
Nomenclature	• Human readable • Backwards compatible (inheritability) • Automatizable
Identification	• Accessible • Fast, user friendly • Accurate

How do we compare bacterial genomes?
Genomic data analysis strategies



De-novo assembly

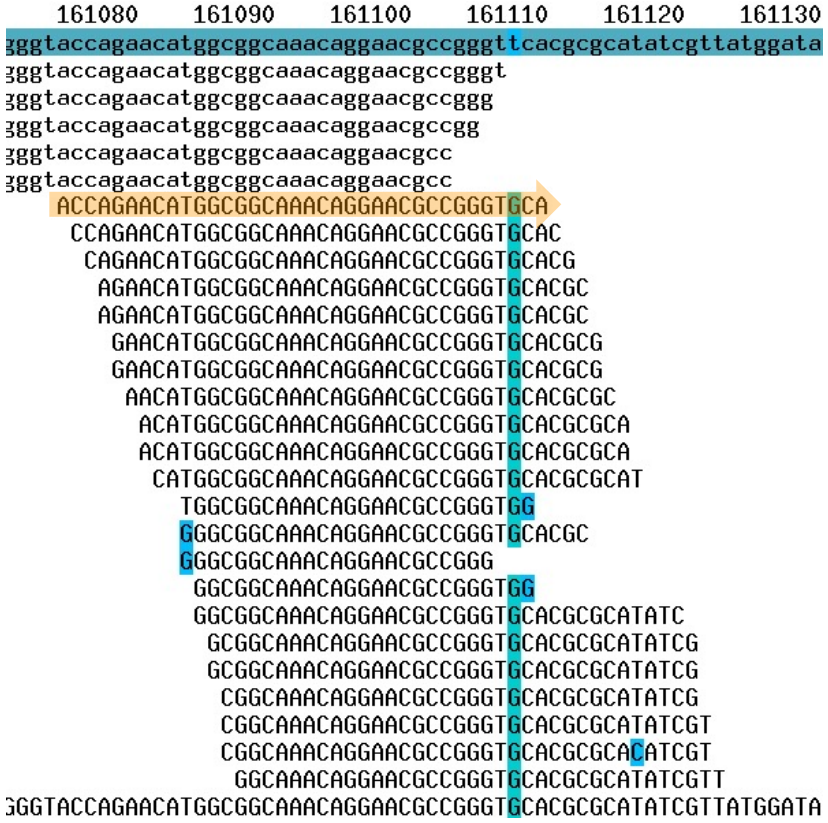




ST	gapA	infB	mdh	pgi	phoE	rpoB	tonB
14	1	6	1	1	1	1	1
15	1	1	1	1	1	1	1
265	3	1	1	1	1	1	1

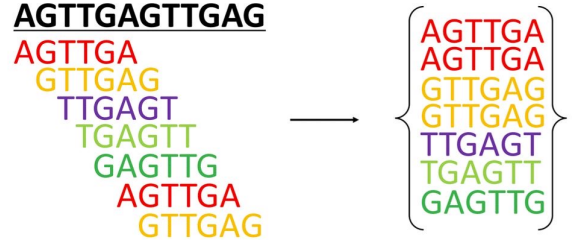
Gene-by-gene approaches (MLST)

Mapping



Single Nucleotide polymorphisms (SNPs)

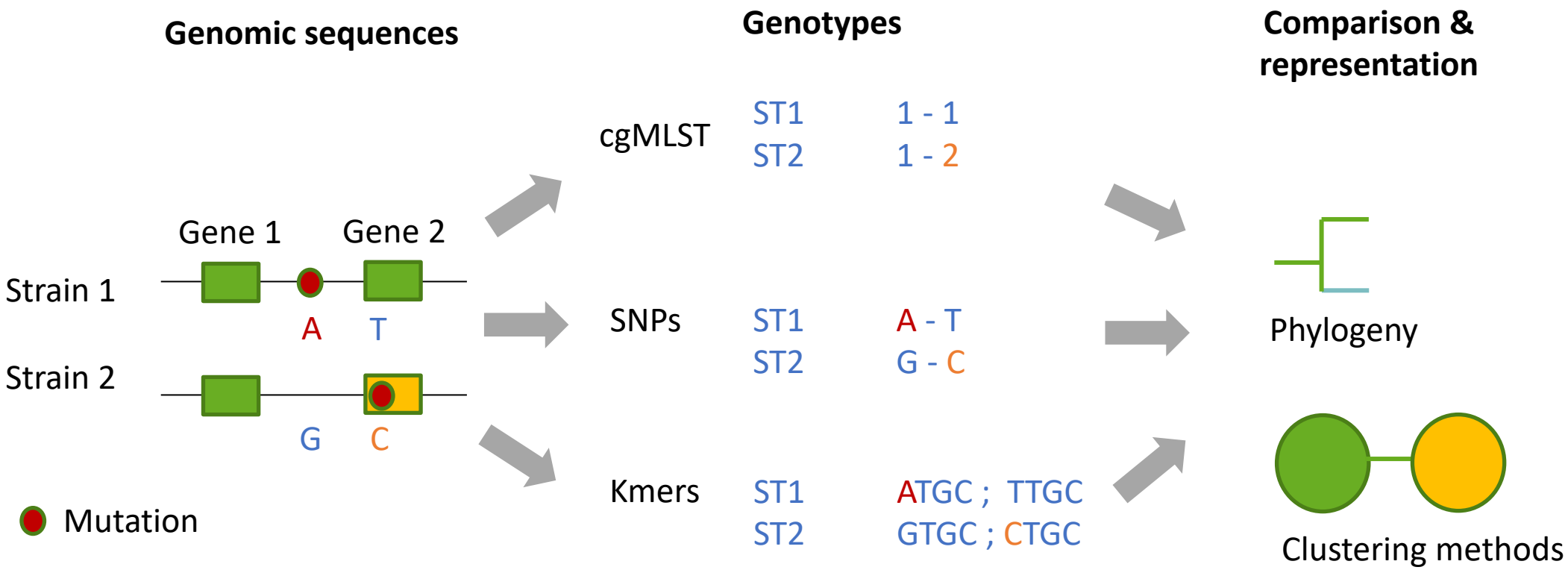
K-mers



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

K-mer approaches

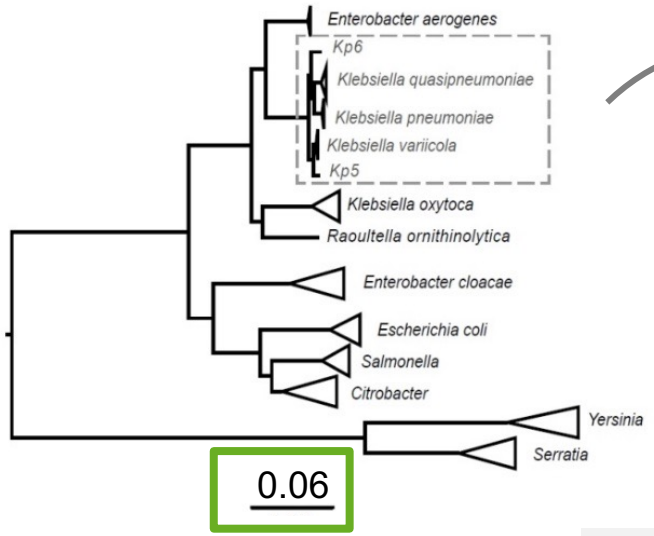
Comparative genomics: 3 approaches



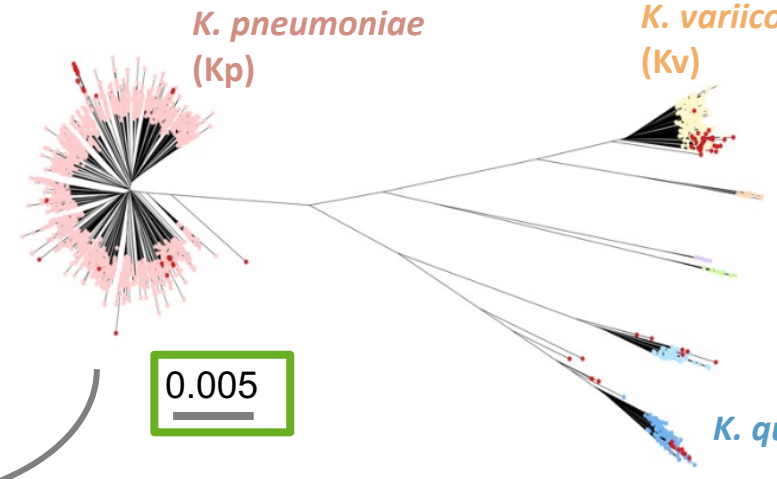
Proposed strain taxonomy systems:

- cgMLST types & multilevel classifications: hierarchical clustering (EnteroBase), LIN codes (BIGSdb)
- SNP address, phylogenetic nomenclatures, *e.g.* Typhi 3.2.4.1
- K-mer based methods (PopPUNK)

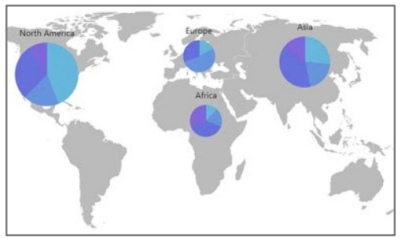
The multiple levels of phylogenetic structure

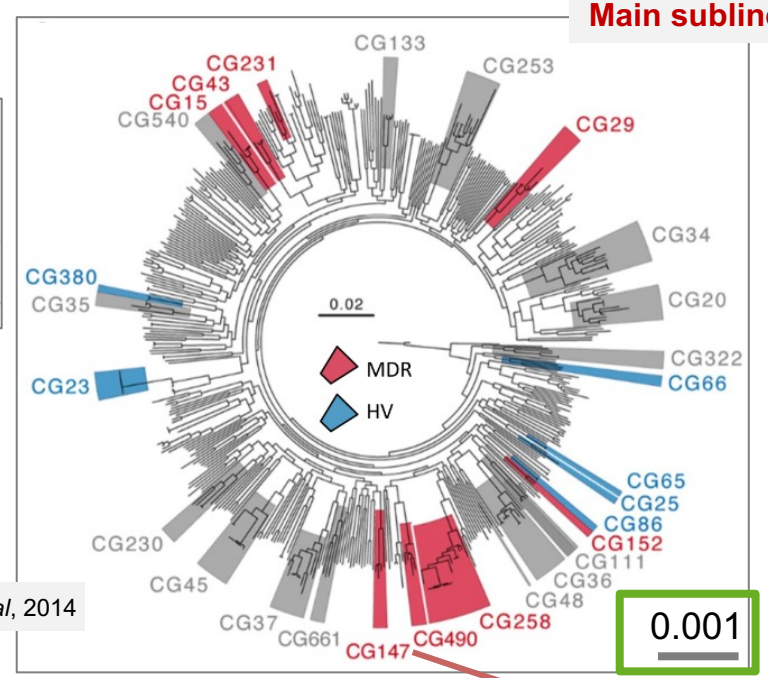


KpSC



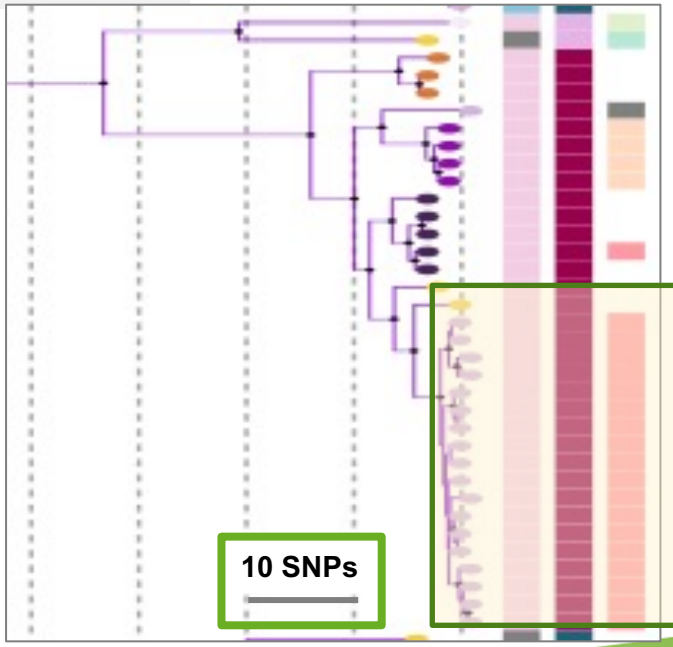
Main sublineages





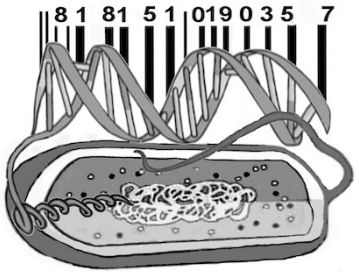
Holt *et al*, 2014

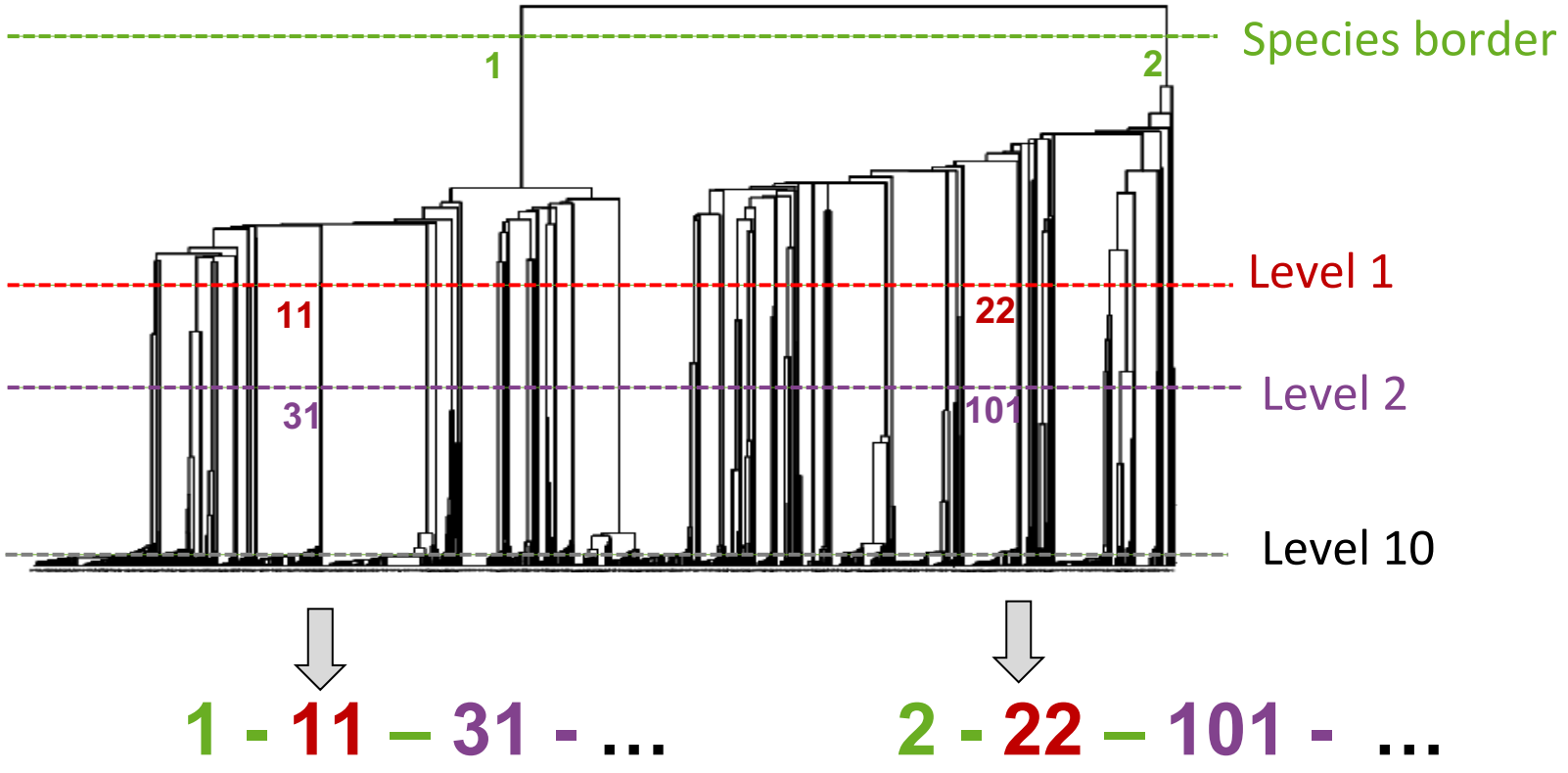
Emerging 'strains'



Strain A

Multilevel classification

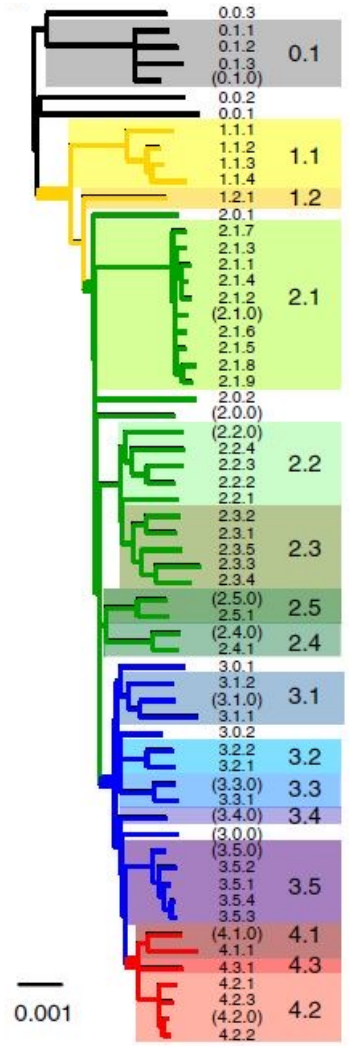




Unique ‘barcode’ for each strain,
capturing group appartenance at various phylogenetic depths

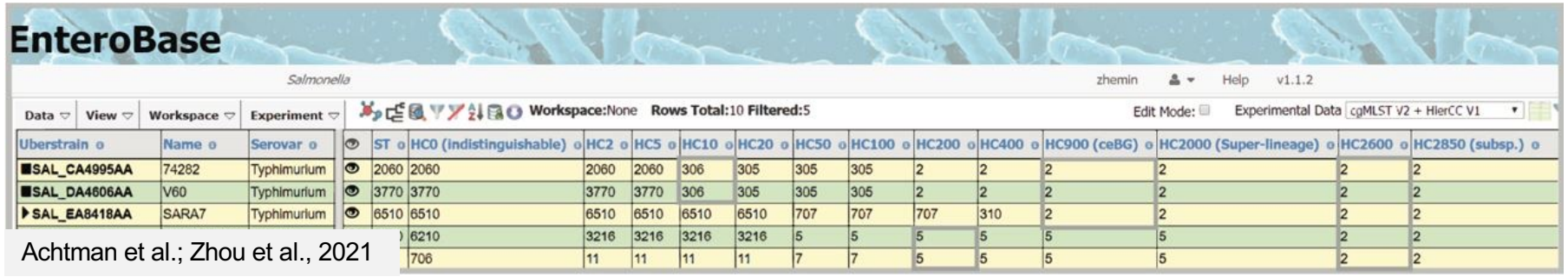
Multilevel genomic classifications

Phylogenetic nomenclatures



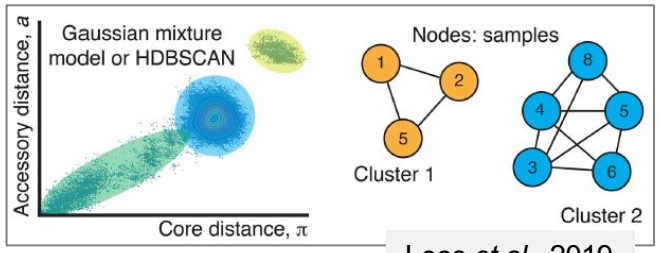
Wong *et al*, 2015; Coll et al., 2014

Hierarchical clustering



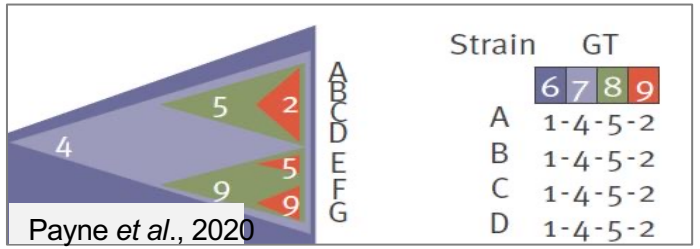
Achtman et al.; Zhou et al., 2021

PopPUNK



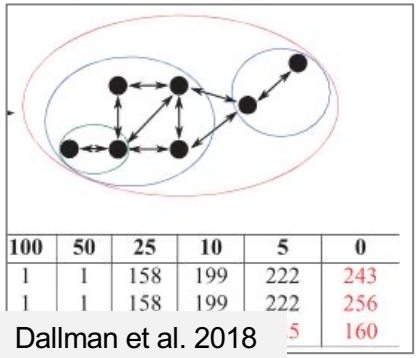
Lees *et al*., 2019

Multilevel genome typing



Payne *et al*., 2020

'SNP address'



100	50	25	10	5	0
1	1	158	199	222	243
1	1	158	199	222	256
				5	160

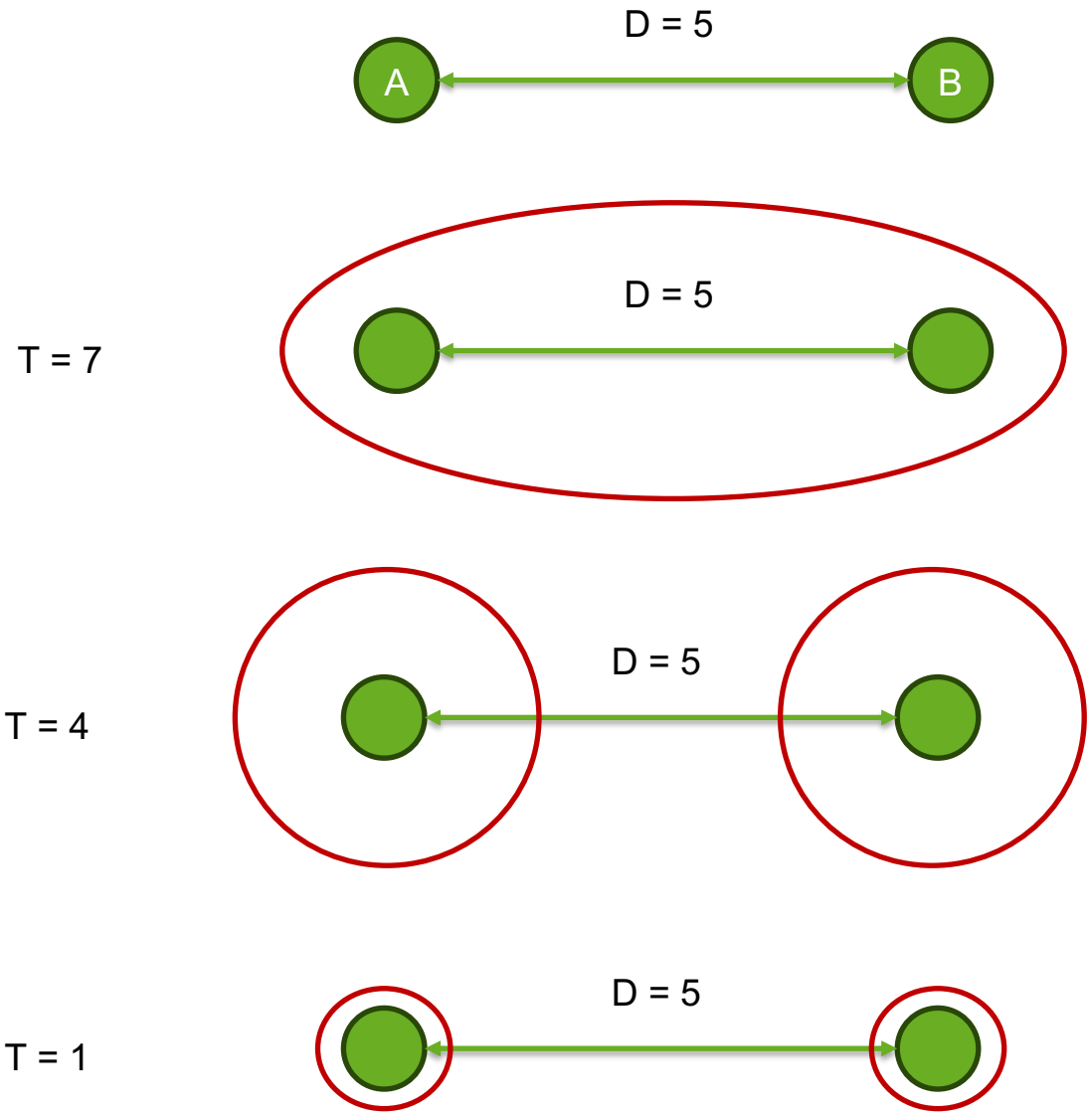
Dallman et al. 2018

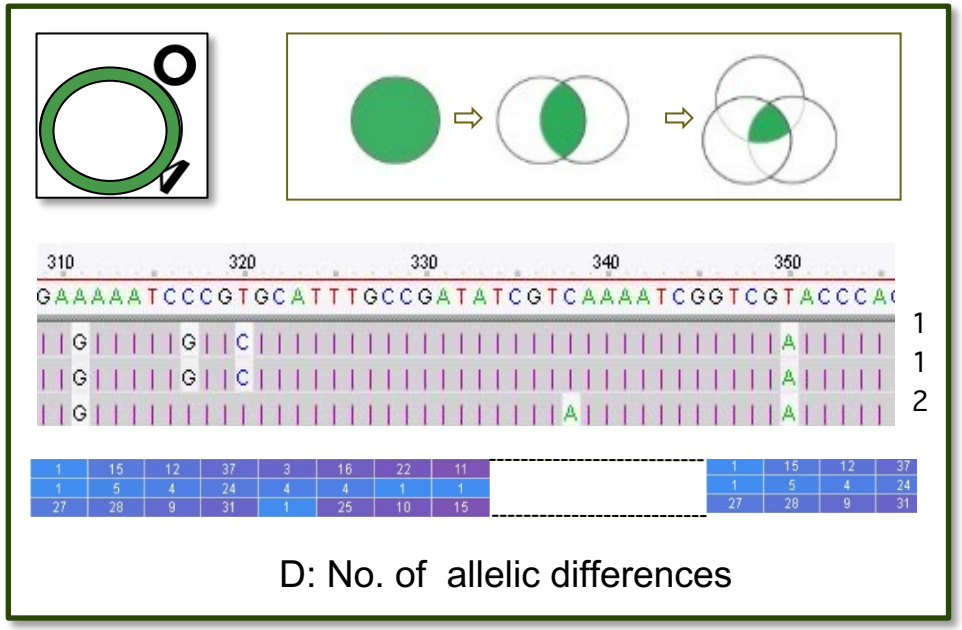
LIN codes (Life Identification Numbers)

% identity threshold	60	70	75	80	85	90	95	98	98.5	99	99.25	99.5	99.75	99.9	99.925	99.95	99.975	99.99	99.999	99.9999
Position	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
G1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Vinatzer *et al*., 2017

Classification using cgMLST allelic differences

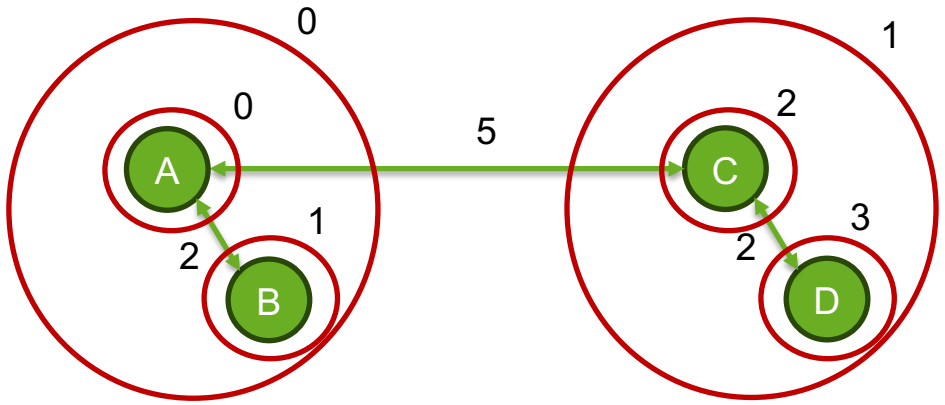


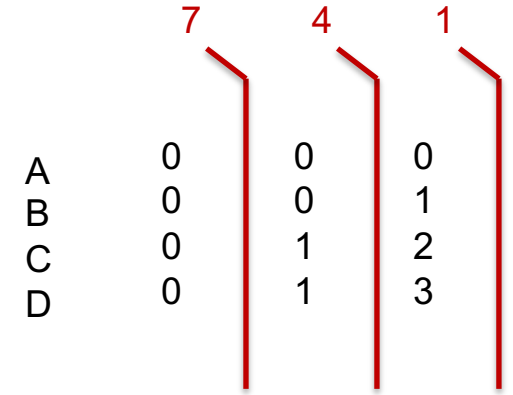


Thresholds

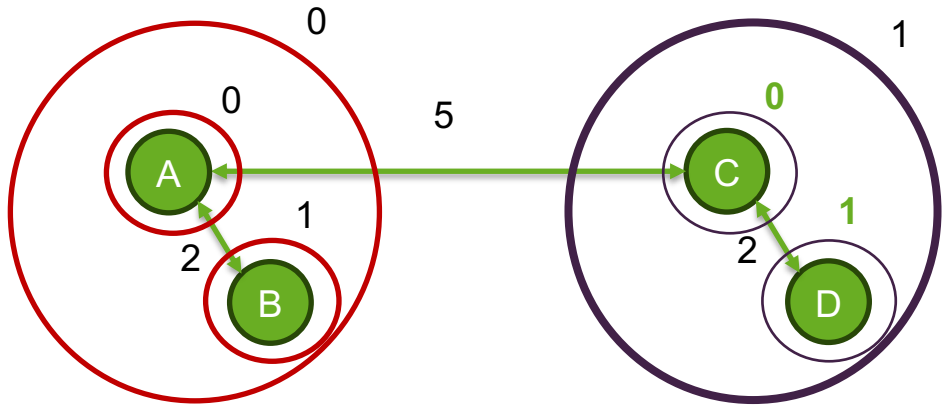
Genomes		7	4	1
	A	0	0	0
	B	0	1	1

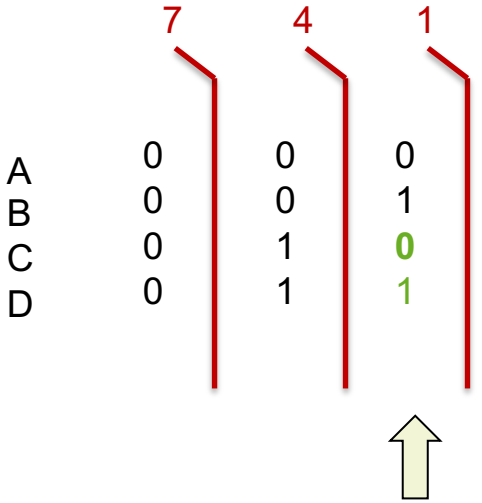
Classification using cgMLST allelic differences





Independent numbering across levels



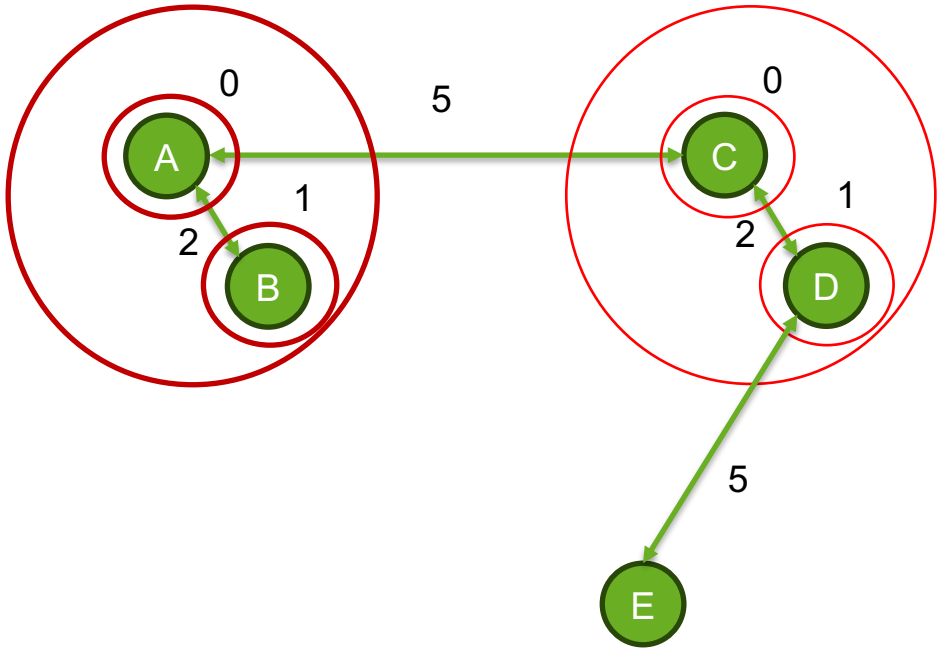


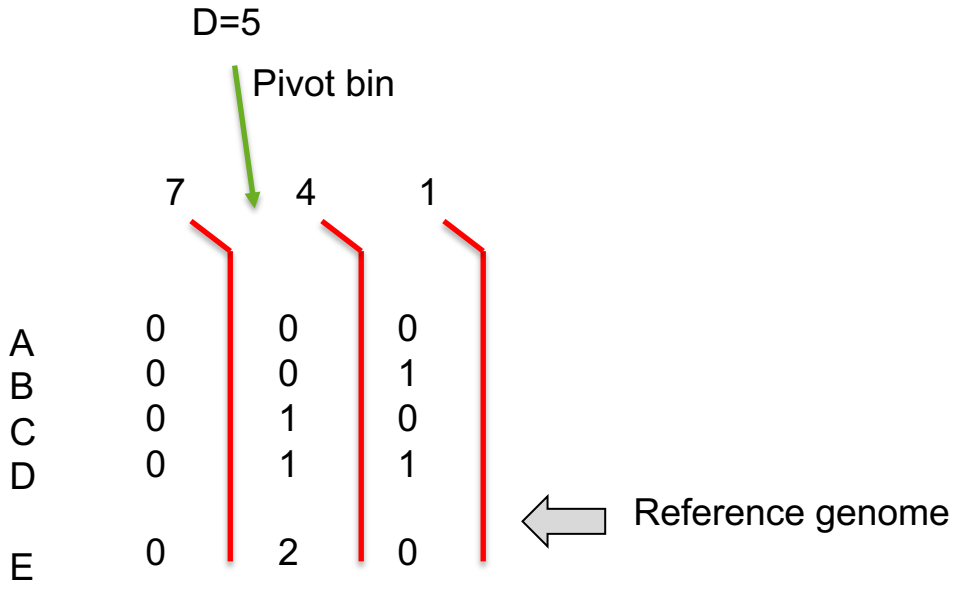
Subdivisions of prefix 0_1

Interdependent numbering across levels

- avoids redundancy
- Keeps identifiers small

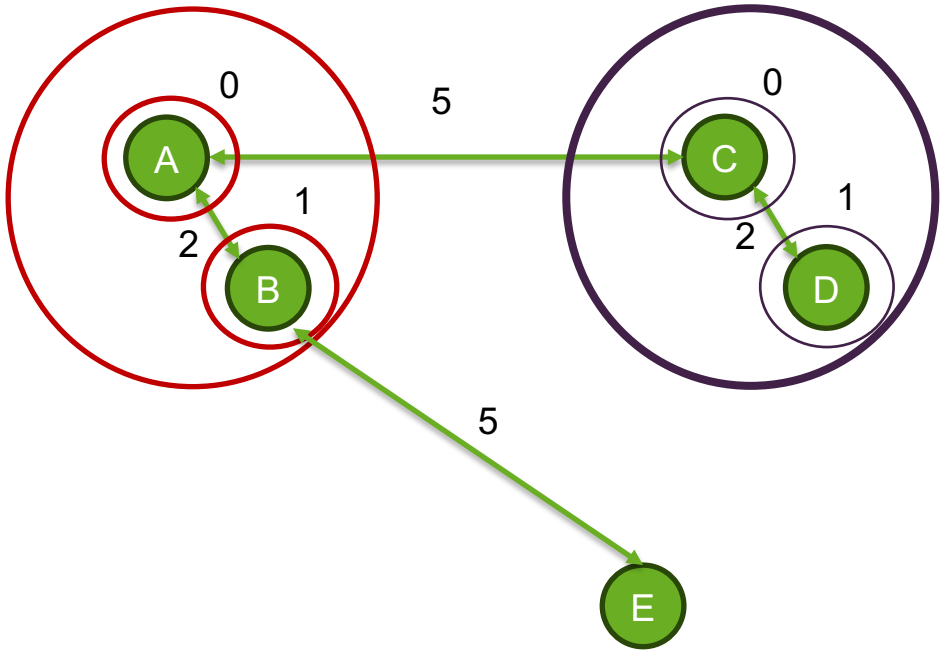
Algorithm for inter-level dependency

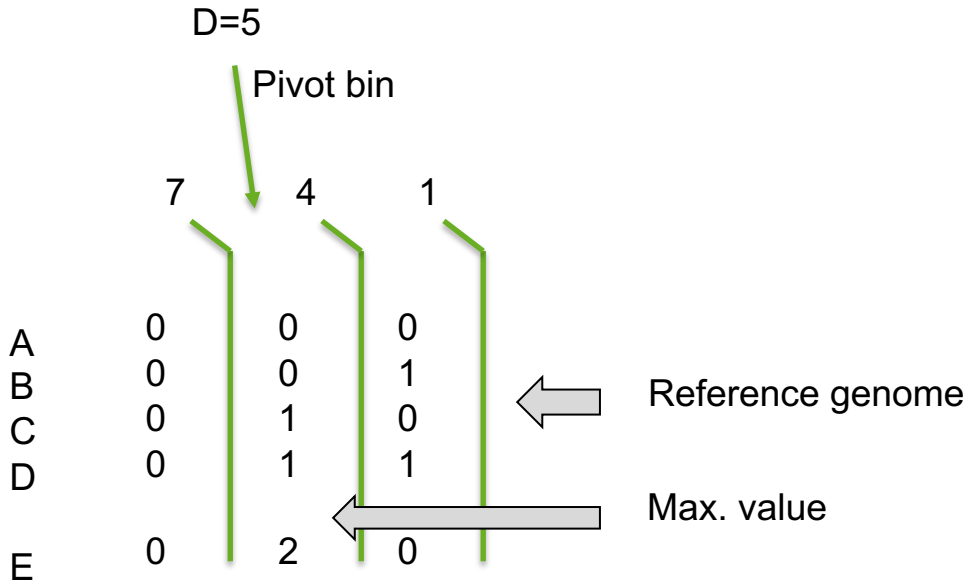




(i) Copy prefix up to pivot (excluded)
(ii) Increment max value in pivot by 1
(iii) Initialize downstream levels with 0

Algorithm for inter-level dependency



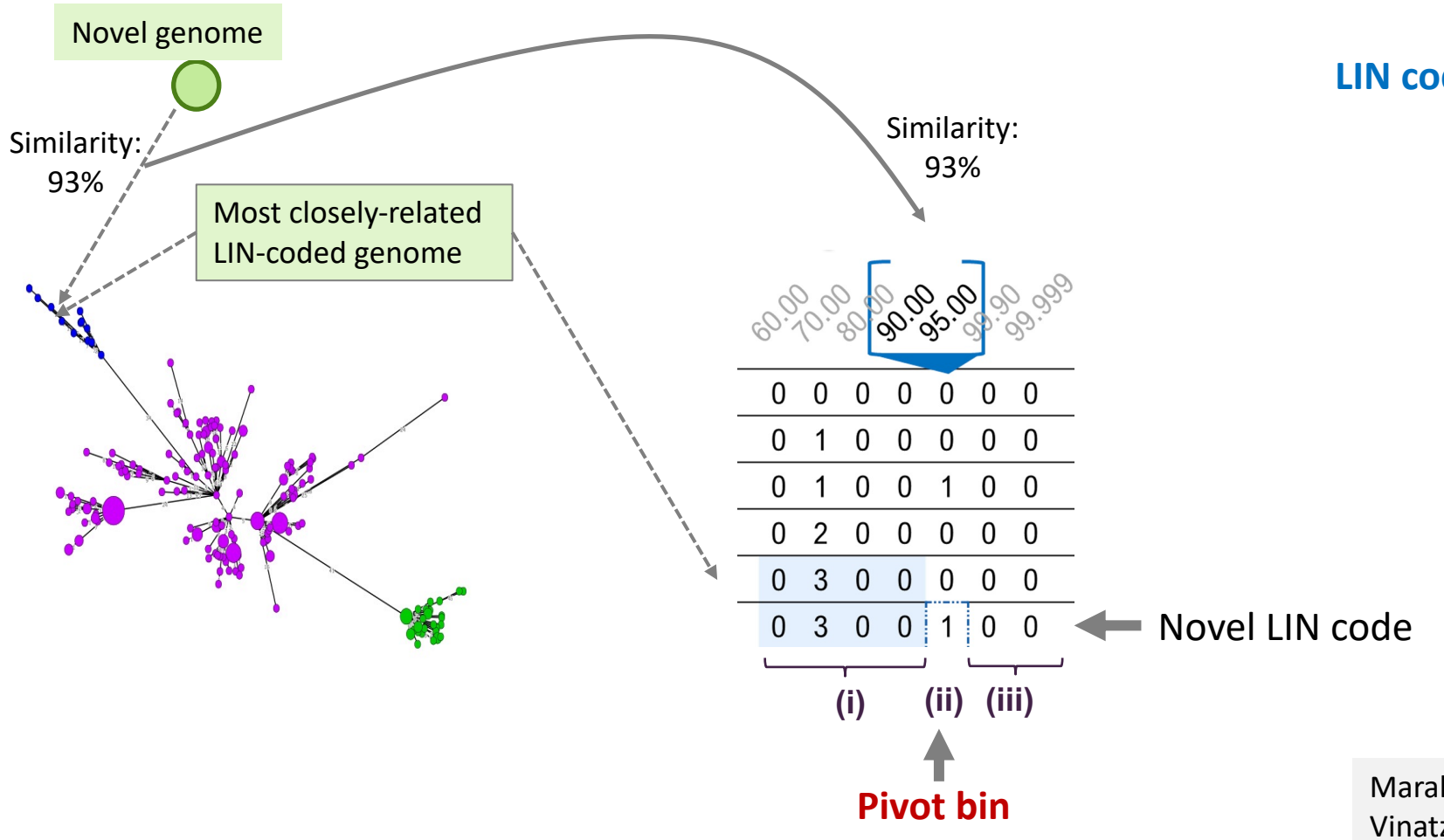


(i) Copy prefix up to pivot (excluded)
(ii) Increment max value in pivot by 1, among all codes that share the same prefix
(iii) Initialize downstream levels with 0

LIN codes: Life Identification Numbers

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:

« bin where the similarity value falls »

2. Create novel LIN code:

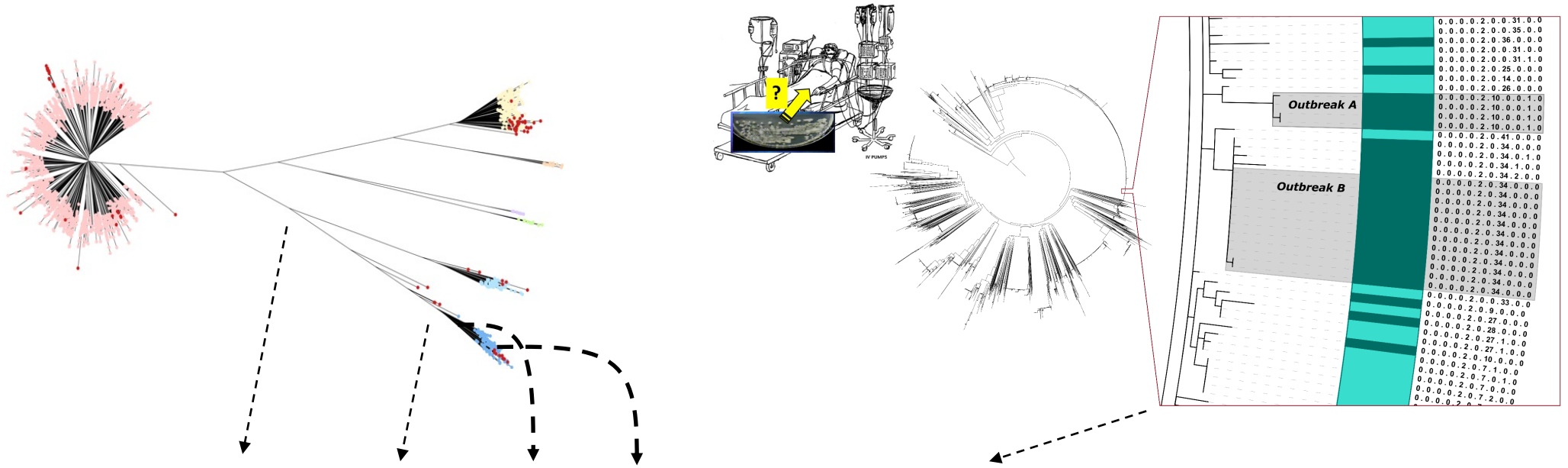
(i) Prefix up to pivot bin, excl.
(ii) +1 in pivot bin
(iii) 0 in downstream bins

Marakeby et al., PLoS One, 2014
Vinatzer *et al.,* 2017, Antonie van Leeuwenhoek

Hennart *et al.,* Mol Biol Evol, 2022 (cgMLST-based)

➡ Stable by design: attributed codes are fixed

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

cgMLST LIN codes for KpSC genomes: How to understand thresholds

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

	Bin number:	1	2	3	4	5	6	7	8	9	10
Left bin threshold	Max. allelic difference	629	610	585	190	43	10	7	4	2	1
Right bin threshold	Min. allelic difference	610	585	190	43	10	7	4	2	1	0
Right bin threshold	Max. allelic *similarity*	19	44	439	586	619	622	625	627	628	629
Left bin threshold	Min. allelic *similarity*	0	19	44	439	586	619	622	625	627	628

Missing data: tolerated, and ignored in comparisons

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

Bin number:	1	2	3	4	5	6	7	8	9	10	
allelic differences *	629	610	585	190	43	10	7	**4**	**2**	**1**	
similarity %	0	3.02	6.99	69.79	93.16	98.41	98.88	99.36	99.68	99.84	100

* Left bin threshold, Defined based on 629 comparisons

If 1 missing allele out of 629, and 1 mismatch:
$d = 1/628 > 1/629$

- Use relative similarity (ratio)
- Denominator = No. called alleles in both

	4	**2**	**1**
A	0	0	0
B	0	0	1
C	0	1	0

Three loci out of 629 (all others called and identical)

A	15	2	1
B	15	2	1
C	24	-	1

$s(A\text{-}C)=0.998407 < 0.998410$ - : missing (uncalled)

Missing data & coincident cgSTs

	Locus 1	Locus 2	Locus 3	…	Locus L
cgST1	X	1	1	…	1
cgST2	2	X	1	…	1
excluded	3	X	X	…	X
Strain A	2	1	1		1

LIN code 1 (*d*=0)

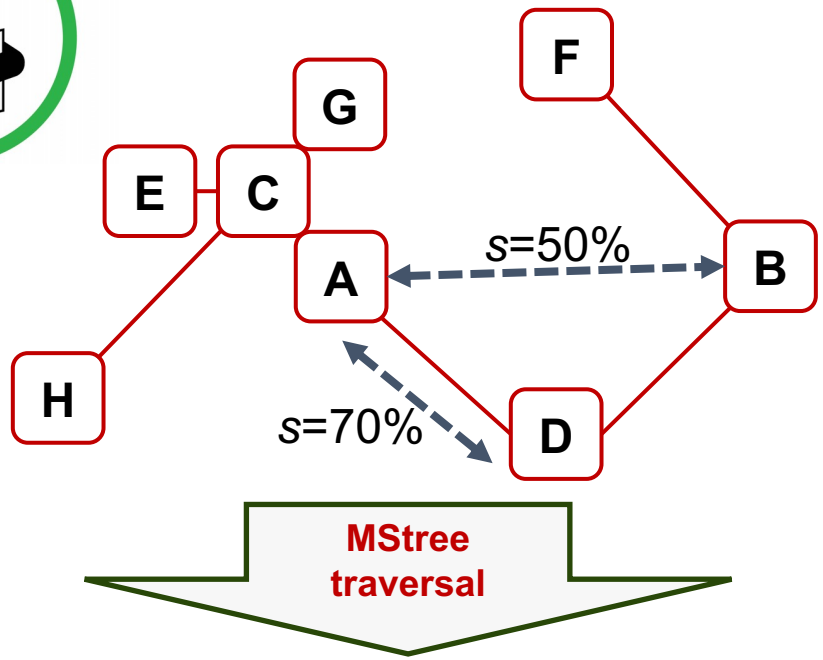
No cgST; no LIN code (*d*>tt)

Multiple matches

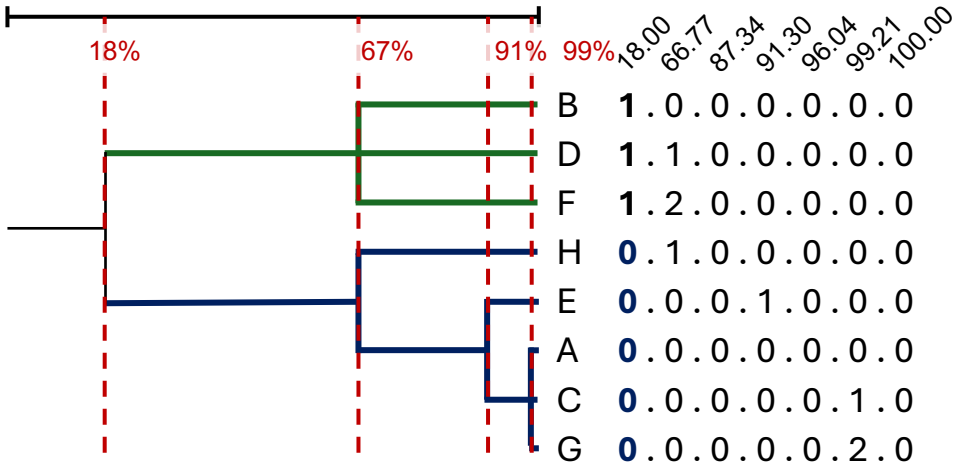
cgST1; cgST2; LIN code 1

- « coincident cgSTs »: differ only by missing data pattern; have *d*=0; same LIN code
- A genomic sequence can match several cgSTs due to the existence of missing data
  … but has a unique LIN code
- Core genome Sequence Types (cgST) are defined only if Number missing data < tolerance threshold (tt); and are then LIN-encoded

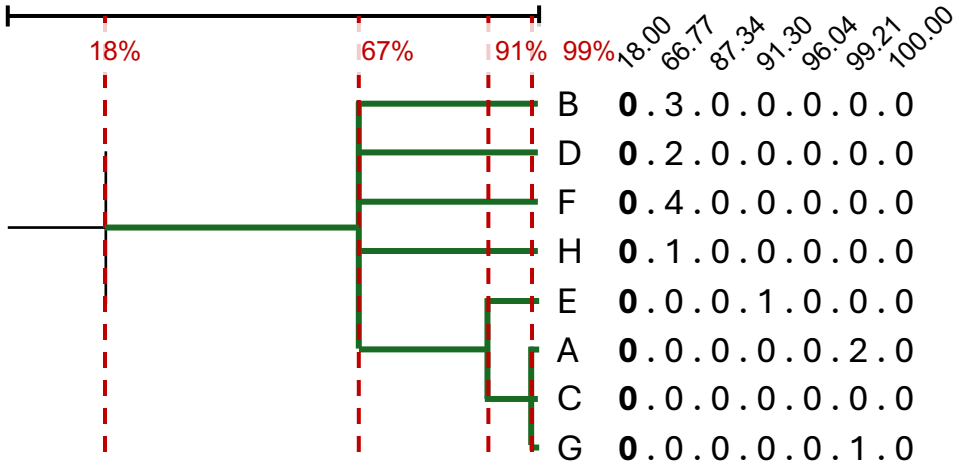
LIN codes & input order (fragmentation)



Input Order: A → B → C → D → E → F → G → H



Input Order: C → G → A → E → H → D → B → F



Solution: use Minimum spanning tree traversal as input order

- ✓ Minimizes partitions
- ✓ Reproducible classification
- ✓ Need to « LIN code » in batches (implemented in BIGSdb)

LIN codes convey similarity information

***LIN code prefix**: any bin subset that starts from leftmost position of a LIN code*

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
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

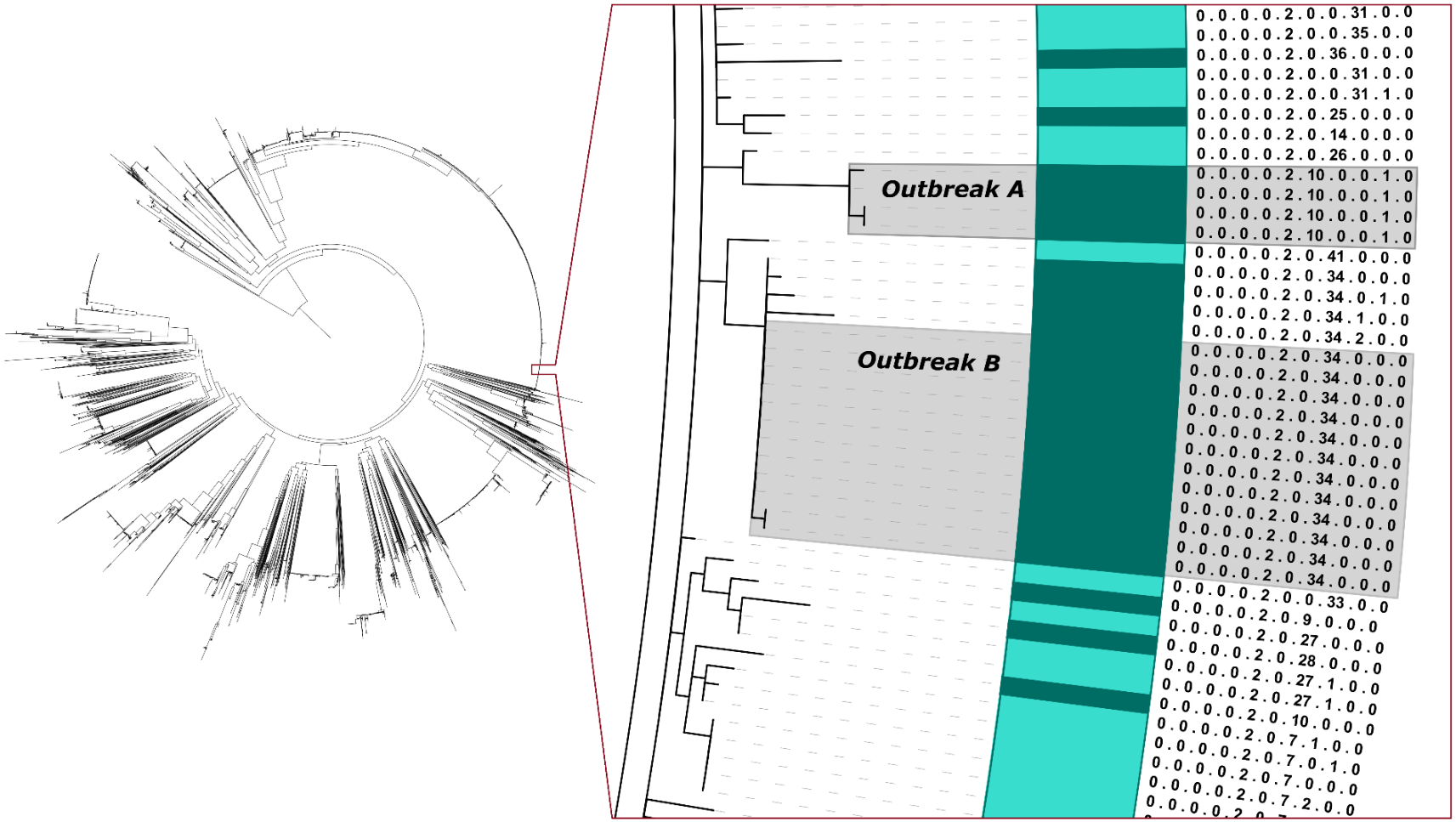
* Right bin thresholds, excluded

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

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

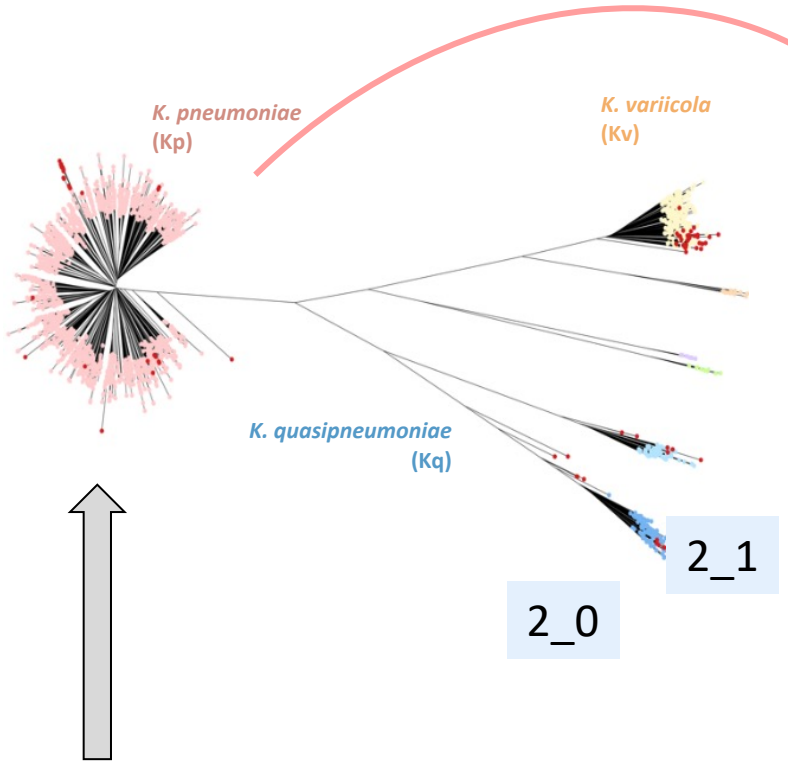
The shared prefix defines a similarity range: the range of the pivot bin

LIN codes can be used to track outbreak strains

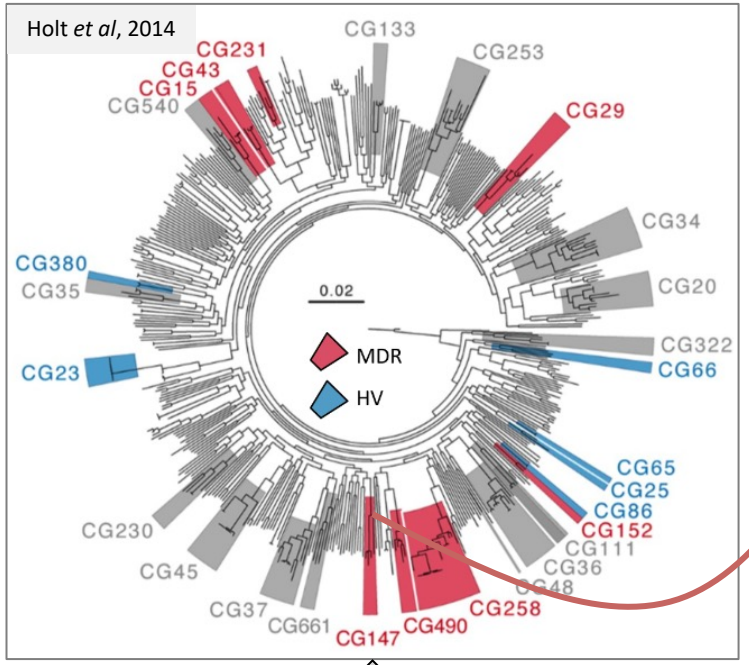


- Only the almost identical genomes may have same LIN code
- For *K. pneumoniae*, the 629-loci cgMLST scheme has discrimination limitations

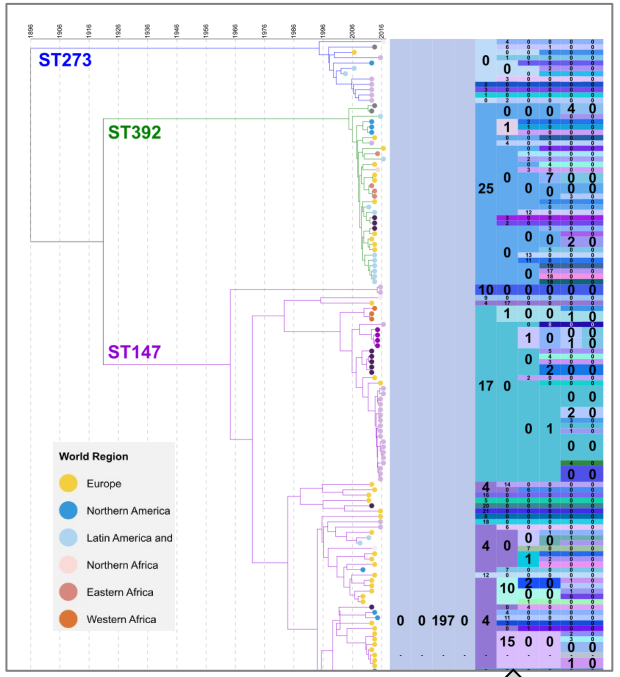
LIN codes prefixes are diagnostic markers



0_0 (*K. pneumoniae*)



0_0_197_0 (CG147)



0_0_197_0_4_15_0_0
('4_15 subclone')

- Prefixes mark important species, sublineages, clonal groups and finer subdivisions
- Prefixes can be used to label such groups, at any level

LIN code prefixes can get human-readable nicknames

Novel system

LIN prefix	Main ST	Nickname
0_0_105_6	258	Alpha
0_0_105_0	340	Beta
0_0_105_2	11	Gamma
0_0_105_11	11	Delta
0_0_105_1	437	Epsilon
0_0_105_29	11	Theta
0_0_105_7	895	Iota

Inherit from serotypes

LIN prefix	Main serotype	Nickname
0_0_1_0	Typhimurium	Typhimurium
0_0_1_6	Enteritidis	Enteritidis
0_0_1_2	Newport	Newport
0_0_1_11	Typhi	Typhi
0_0_1_1	Paratyphi B	Paratyphi B
0_0_1_29	Choleraesuis	Choleraesuis
0_0_1_7	Agona	Agona

Inherit from MLST

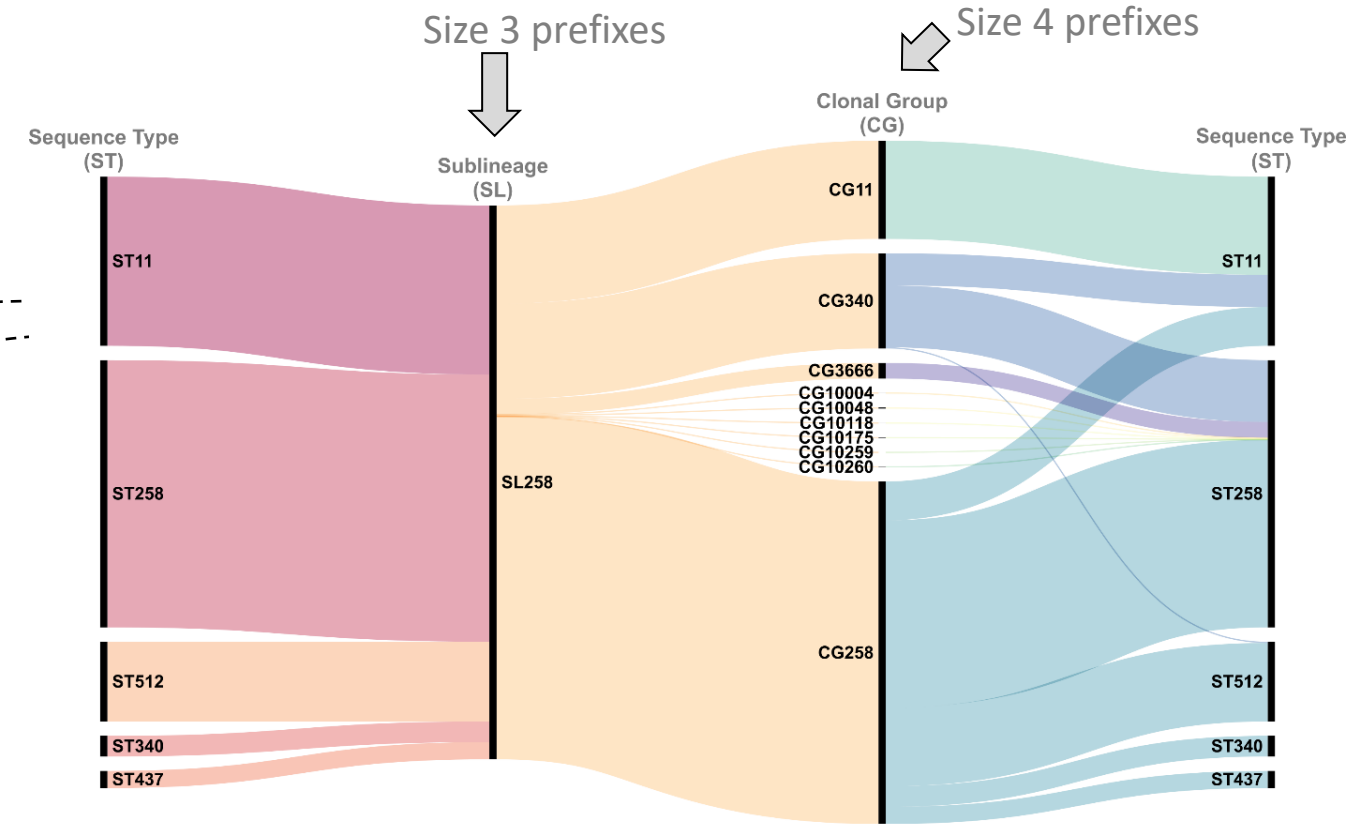
LIN prefix	Main ST	Nickname
0_0_105_6	**258**	**CG258**
0_0_105_0	**340**	**CG340**
0_0_105_2	**11**	**CG11**
0_0_105_11	11	CG3666
0_0_105_1	437	GC10268
0_0_105_29	11	CG12811
0_0_105_7	**895**	**CG895**

(*mock example*)

- LIN code nomenclature: use prefix nicknames to name groups
- Can inherit nickname from previous nomenclature: backwards compatible

Inheritance of MLST ST identifiers on LIN code nomenclature

KpSC LIN codes:
Nicknaming prefixes at levels 3 and 4



Inheritance from 7-gene MLST

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

Don't say ST11, say CG11 or CG258; or SL258

Nicknames: saying goodbye politely to MLST

LIN prefix	(sub)Species
0_0	*K. pneumoniae*
1_0	*K. variicola* subsp. *variicola*
1_1	*K. variicola* subsp. *tropica*
2_0	*K. quasipneumoniae* subsp. *quasipneumoniae*
2_1	*K. quasipneumoniae* subsp. *similipneumoniae*
3_0	*K. quasivariicola*
4_0	*K. africana*

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 nomenclature, inherited from MLST
- MLST still maintained in parallel (expansion of familiar system)

Klebsiella pneumoniae identification based on LIN codes

https://bigsdb.pasteur.fr
https://pubmlst.org

kjolley / BIGSdb

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

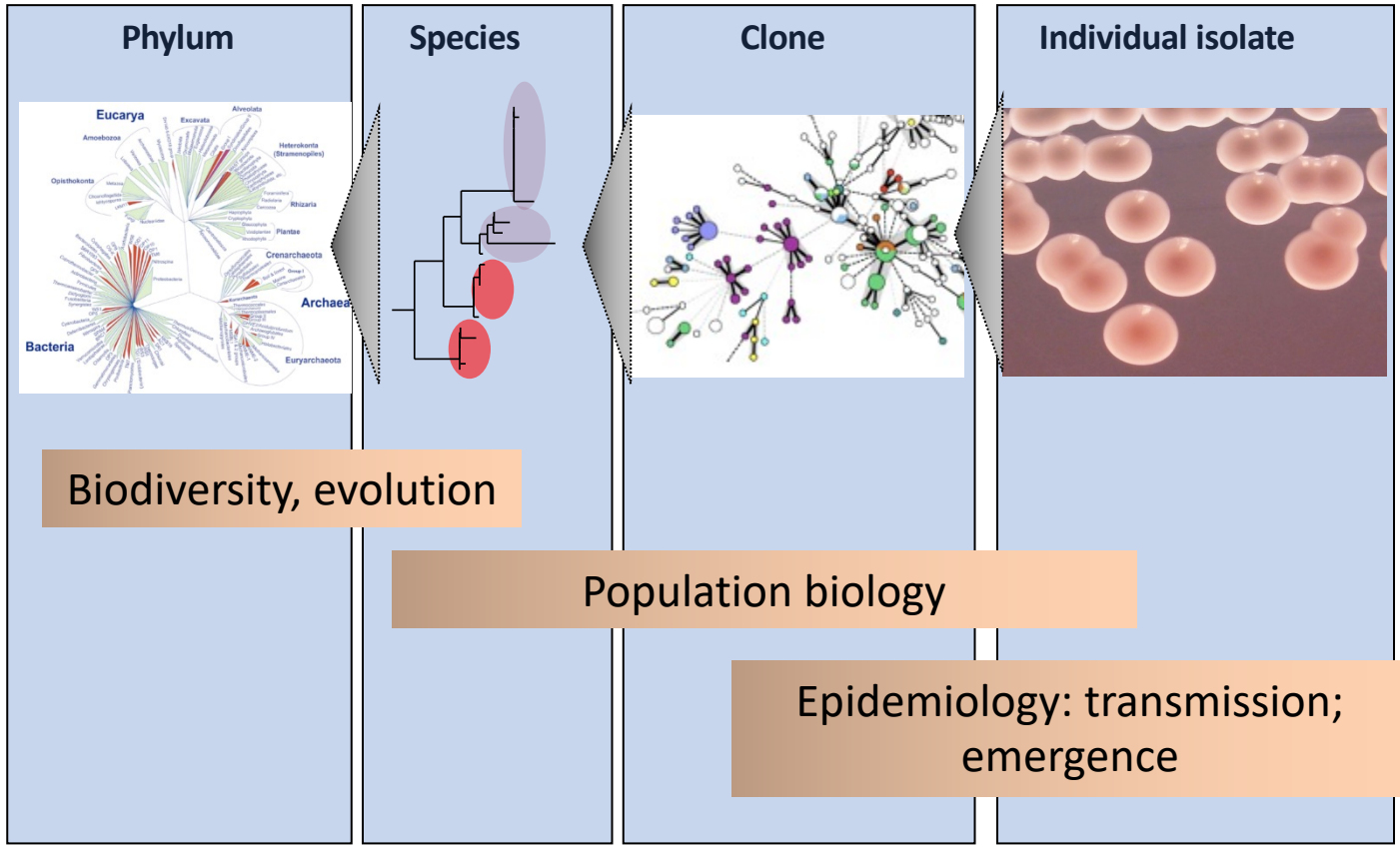
0_0_105

SL258

LIN code taxonomy

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

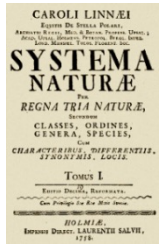
LIN codes strain taxonomy

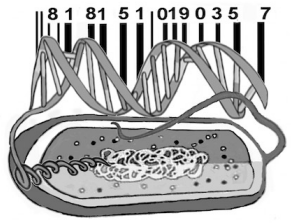


- ✓ Stable classification
- ✓ Multi-level
- ✓ Automated
- ✓ Human readability
- ✓ Backward compatibility
- ✓ User-friendly genome identification

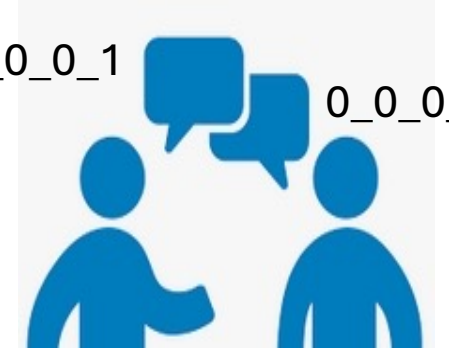
Linnaean taxonomy

Genomic taxonomy



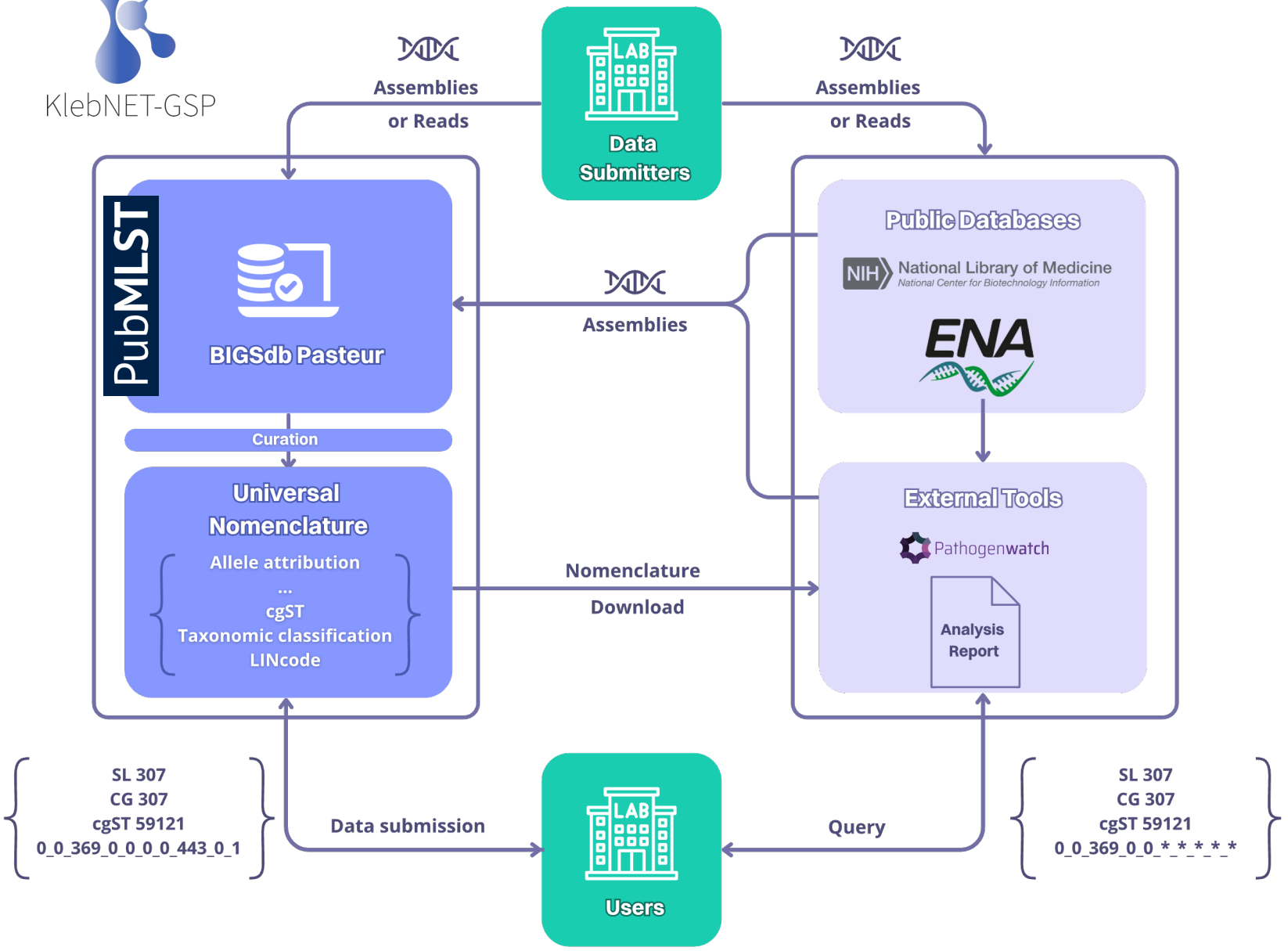


Can communicate on subtypes, without sequence data exchange



LIN code taxonomy ecosystem

KlebNET-GSP



- Free data submission; Private spaces (embargo)
- SRA included via Pathogenwatch
- Nomenclatures available for download (APIs)
- External usage possible, but may attribute incomplete codes
- Need to submit to source database to define novel codes
- *K. pneumoniae, S. pneumoniae, S. aureus, Campylobacter, Neisseria, Listeria…*

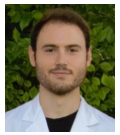
Contributors

Biodiversity & Epidemiology of Bacterial Pathogens


Virginie PASSET


Carla RODRIGUES


Sebastien BRIDEL


Jose Francisco DELGADO BLAS


Mélanie HENNART

Martin RETHORET-PASTY
Audrey COMBARY


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

Further reading

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

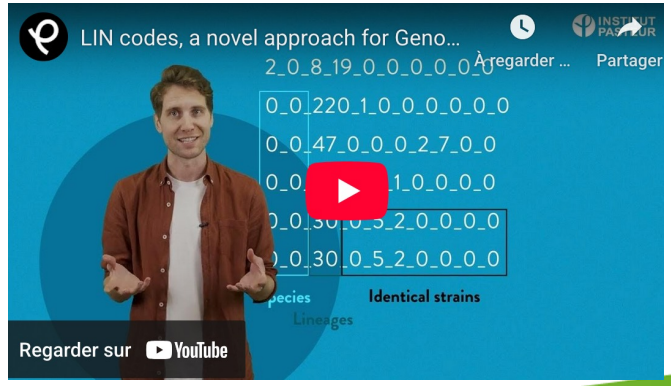
Federica Palma, Melanie Hennart, Keith A. Jolley, Chiara Crestani, Kelly L. Wyres, Sebastien Bridel, Corin A. Yeats, Bryan Brancotte, Brice Raffestin, Sophia David, Margaret M. C. Lam, Radosław Izdebski, Virginie Passet, Carla Rodrigues, Martin Rethoret-Pasty, Martin C. J. Maiden, David M. Aanensen,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

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

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

LIN codes in BIGSdb-Pasteur *Klebsiella pneumoniae*:

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



Acknowledgements

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