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

Federica Palma[1,2,#], Melanie Hennart[1,#], Keith A. Jolley[3], Chiara Crestani[1], Kelly L. Wyres[4,5], Sebastien Bridel[1], Corin A. Yeats[6], Bryan Brancotte[7], Brice Raffestin[8], Sophia David[6], Margaret M. C. Lam[4], Radosław Izdebski[9], Virginie Passet[1], Carla Rodrigues[1], Martin Rethoret-Pasty[1], Martin C. J. Maiden[3], David M. Aanensen[6], Kathryn E. Holt[4,10], Alexis Criscuolo[2], Sylvain Brisse[1,2,*]

[#] Equal contribution

Affiliations

[1] Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France

[2] Institut Pasteur, Université Paris Cité, Biological Resource Center of Institut Pasteur, Paris, France

[3] Department of Biology, University of Oxford, Oxford, United Kingdom

[4] Department of Infectious Diseases, School of Translational Medicine, Monash University, Melbourne, Australia

[5] Centre to Impact AMR, Monash University, Clayton, 3800, Australia

[6] Centre for Genomic Pathogen Surveillance, Pandemic Sciences Institute, University of Oxford, Oxford, United Kingdom

[7] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France

[8] Institut Pasteur, Université Paris Cité, HPC Core Facility, Paris, France

[9] Department of Molecular Microbiology, National Medicines Institute, Warsaw, Poland

[10] Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom

[*]**Correspondence:** Prof Sylvain Brisse, Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, 25-28 rue du Docteur Roux, F-75724, Paris, France; Phone: +33 1 45 68 83 34; E-mail: sylvain.brisse@pasteur.fr

Abstract

Unified strain taxonomies are crucial for fostering international communication in microbiological research and for the epidemiological surveillance of bacterial pathogens. While multilocus sequence typing (MLST) has served as a foundation of strain taxonomy for two decades, whole genome sequencing enables more precise classifications and significantly improves discriminatory resolution. The core genome-wide extension of MLST (known as cgMLST) thus holds great promise for strain genotyping and classification, but its implementation faces challenges that include missing data, potential instability of cluster-based nomenclatures, and the necessity to ensure backwards compatibility with MLST identifiers. Life Identification Number (LIN) codes offer a solution by providing multi-level classification groups that are inherently stable. Here, we present, consolidate, and extend the cgMLST-based LIN code approach. We first develop a nicknaming system for LIN code prefixes, which enables flexible human-readable strain nomenclatures. Using *Klebsiella pneumoniae* (Kp) as an example, LIN code nicknames were attributed by inheritance from MLST identifiers, thus perpetuating the legacy of MLST nomenclatures in the genomic era. We show that while 7-gene MLST sometimes conflates unrelated sublineages into the same ST, cgMLST-based LIN codes are highly concordant with phylogenetic relationships. We implement this novel LIN code-based nomenclature in the BIGSdb platform, and illustrate, with Pathogenwatch, how it can also be used in other genomic epidemiology platforms. Finally, we demonstrate the value of LIN codes for tracking the strain diversity within high-risk internationally disseminated clonal groups of Kp and protracted outbreaks. Given its stability, precision, and flexibility, we recommend the adoption of the cgMLST-based LIN code taxonomic approach for Kp and suggest that this approach is widely applicable to other bacterial pathogens.

Introduction

Taxonomies of bacterial strains responsible for infectious diseases are essential resources to ensure effective communication in population biology, epidemiological surveillance, and public health response to outbreaks. As illustrated by the SARS-CoV-2 variant nomenclature system, simple nicknames (*e.g.*, Alpha, Delta, Omicron) for pathogen variants can greatly improve communication between different public health sectors (Konings et al., 2021; Rambaut et al., 2020).

Currently, there are neither classification nor nomenclature standards to define sublineages, variants, types or clones (hereafter, collectively called "strains") within bacterial species ("International Code of Nomenclature of Prokaryotes," 2019). Ad-hoc phenotypic (*e.g.*, serotypes) and genotypic (*e.g.*, sequence types) approaches have long been used to define strains of particular species, but the advent of universally applicable whole genome sequencing (WGS) has the potential to refine and generalize strain taxonomy by providing the maximal discrimination needed for epidemiological surveillance, and a harmonized general approach across pathogen phyla (Maiden et al., 2013; Nadon et al., 2017; Struelens and Brisse, 2013). However, few attempts have been made to devise genomic taxonomies and evaluate their general applicability. With WGS implemented worldwide and in all sectors of microbiology (medical, veterinary, food, environmental), a precise and universal approach for describing strains of bacterial species becomes a key need to translate WGS data into relevant information that would support epidemiological surveillance, outbreak investigations, cross-niche or between host transmission detection, and public health actions that need international and cross-sectoral coordination.

Among the broad range of methods developed for bacterial strain typing and group naming (Struelens et al., 1998; van Belkum et al., 2007), multi-locus sequence typing (MLST), based on the analysis of a few (typically seven) conserved loci, was established over the last two decades as the method of choice for strain taxonomy of most bacterial species (Aanensen and Spratt, 2005; Maiden, 2006; Maiden et al., 1998). This gene-by-gene approach was logically extended to the genome scale, with core genome MLST (cgMLST) schemes encompassing thousands of loci (Bialek-Davenet et al., 2014; Maiden et al., 2013). Whether using the classical or the core genome MLST schemes, the "sequence type" (ST) nomenclature system is highly reproducible, portable, and easy to interpret (Feil, 2004). To recognize deeper phylogenetic associations, cgMLST allele profiles can be grouped at any level of similarity by single-linkage clustering or static aggregation to predefined groups or founder genotypes (Zhou et al., 2021).

A novel system for genome classification was proposed by Vinatzer and colleagues, using multi-position numerical codes attributed to each individual genome (Marakeby et al., 2014; Vinatzer et al., 2017). These codes, called Life Identification Numbers (LINs), were designed to encompass all domains of life in a single taxonomy, based on the Average Nucleotide Identity (ANI) metric (Goris et al., 2007; Konstantinidis and Tiedje, 2005). However, the ANI-based genome similarity is imprecise and non-reproducible for nearly identical strains, which are most often compared through sequences of draft genomes. Leveraging the strengths of both approaches, some of us recently proposed combining cgMLST and LIN codes to design taxonomies of bacterial strains within species (Hennart et al., 2022). The use of cgMLST dissimilarities, rather than ANI-based similarities, provides robustness in estimating small-scale genome relationships, which are efficiently summarized by cgMLST LIN codes (hereafter, LIN codes for short).

In this article, we present further developments of the LIN code approach. We first design a nicknaming approach for LIN codes, which can be used to recognize familiar groups that are important in biological research or epidemiological surveillance. We further show the benefit of inheriting these nicknames from MLST identifiers. We additionally describe practical implementations of LIN codes in the widely used genotyping platforms BIGSdb (Argimón et al., 2021; Jolley et al., 2018). We next illustrate the use and benefits of LIN code strain taxonomy using the *Klebsiella pneumoniae* Species Complex (KpSC), a phenotypically and genetically diverse ubiquitous pathogenic group (Wyres et al., 2020). We show that for this pathogen, classical (7-gene) MLST classifications can be misleading, and that LIN codes can pinpoint these cases and mitigate misclassifications. Lastly, we illustrate the benefit of LIN codes for defining and naming intraspecific groups from epidemiologically important phylogenetic lineages down to outbreak strains in a stable way.

Section 1: LIN codes: definitions and practical implementation

The principle of cgMLST-based LIN codes: an overview

Here we explain in more detail how cgMLST-based LIN codes work, as originally proposed (Hennart et al., 2022), before describing new developments and applications of the system (see **Section 2: Novel developments and examples of applications)**. The core genome Life Identification Number classification code system combines the core genome MLST (cgMLST) approach with Life Identification Numbers (LIN) (Vinatzer et al., 2017). The LIN codes consist of multiple (for example, 10) predefined positions (or bins), each corresponding to a (range of) cgMLST profile similarity value, together representing a partition of the complete range [0%-100%]. From left to right, the positions of the code correspond to decreasing allele mismatch dissimilarity, *i.e.*, increasing similarity. The leftmost bins capture the lowest similarities reflective of deep phylogenetic divisions, whereas the rightmost bins capture the highest similarities. Each bin has a left border threshold (inclusive) that corresponds to a maximum number of pairwise allele differences between profiles and is delimited on the right by the next threshold (exclusive, as the threshold value corresponds to the left threshold of the downstream bin).

While any number of bins (up to the number of loci in the cgMLST scheme) can be chosen, in the case of the *Klebsiella pneumoniae* Species Complex (KpSC) used here as an example, 10 bins were determined to define their LIN codes (Hennart et al., 2022). The first four bins represent the deepest hierarchical levels of relatedness, corresponding to species, subspecies, sublineage and clonal group, respectively (Hennart et al., 2022). The last bins delineate six levels of high-resolution relatedness that might be useful for epidemiological surveillance. KpSC profiles are defined using a 629-loci cgMLST scheme; bins 1 to 4 have as right borders 610, 585, 190 and 43 allele mismatches, respectively, while bins 5 to 10 correspond to thresholds 10, 7, 4, 2, 1 and 0 mismatches, respectively. Thus, the first bin corresponds to the range [629-610[of cgMLST mismatches (the '[' indicates the value 610 is excluded), whereas the last one corresponds to the range [1-0[(note that it excludes complete identity, *i.e.*, 0 mismatch, 629 matches: in this case, the LIN code is simply copied from the reference, see below).

Formally, LIN codes are attributed to core genome Sequence Types (cgST) (Hennart et al., 2022). Therefore, before assigning LIN codes, cgMLST profiles must be assigned to cgSTs. Like the ST designation in classical 7-gene MLST, a cgST is defined for each unique cgMLST profile, characterized by a unique combination of alleles at all loci of the scheme. Profiles with too many missing loci can be filtered out at this stage. In practice, for the KpSC, cgMLST profiles are assigned to a cgST only when they comprise fewer than 30 missing alleles (*i.e.*, equal to or more than 600 called alleles). Profiles with 30 (4.77%) or more missing alleles (which are likely to correspond to poor quality genomes) are not considered further, and therefore not included in the KpSC LIN code taxonomy. For any LIN code taxonomy, the proportion of tolerated missing data for cgST assignment can be set to higher values (to increase the proportion of coded genomes) or lower values (to improve the precision of LIN code classifications).

LIN codes are created for each distinct cgST. The formal process of LIN code assignment from cgMLST data, first proposed in (Hennart et al., 2022), is presented in **Box 1** and summarized in **Figure 1**. The system is initialized by creating, for an initial allelic profile, a LIN code with the integer value 0 at every bin. This initial profile can be chosen randomly or based on a reference genome of the species under consideration, as convenient. The next steps are the same for all subsequent individual cgSTs.

Box 1. The formal process of assigning LIN codes

The LIN code of the first allelic profile is attributed 0 in every bin. Next, each new allele profile *j* is encoded from its closest already encoded profile *i* (*i.e.*, that maximizes the allele similarity percentage s_{ij}). After determining the pivot bin *p*, such that $s_{ij} \in [s_p, s_{p+1}[$ (*i.e.*, right threshold exclusive), the encoding of the new profile *j* is performed in three steps:

(i) the same prefix as code *i* is attributed up to the bin $p-1$ (inclusive);

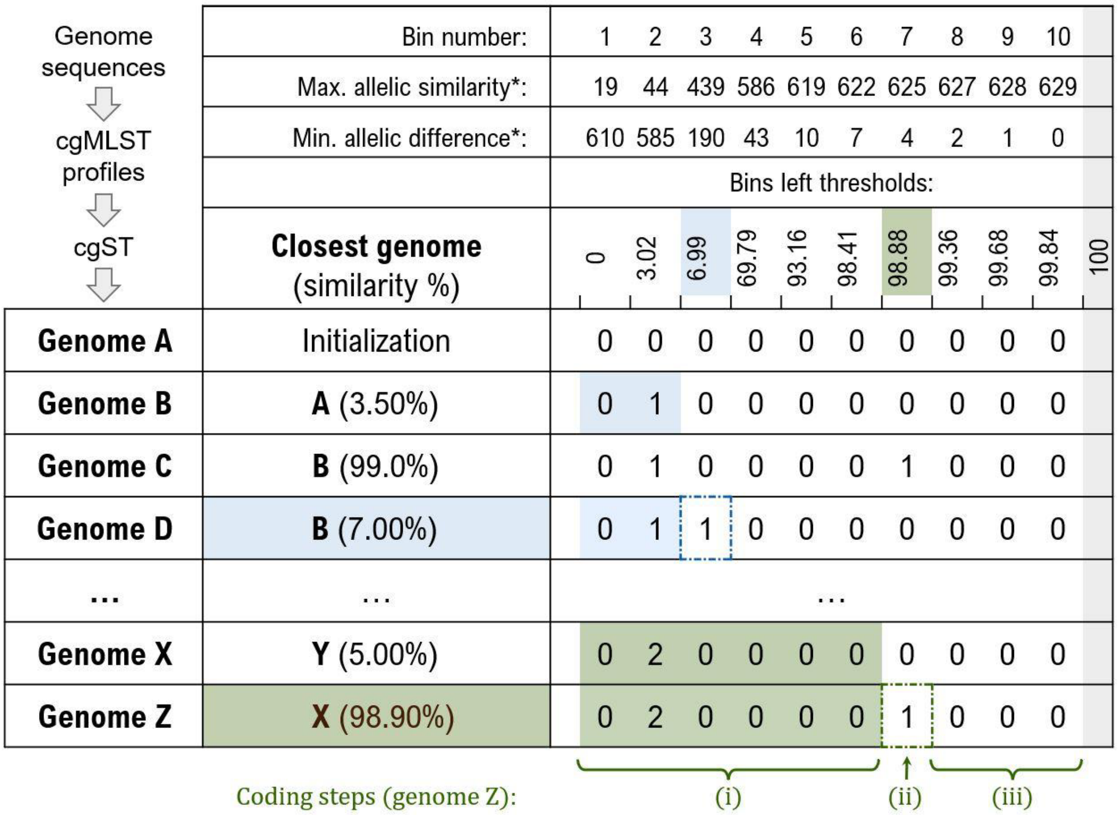
(ii) for the pivot bin *p*: the maximum value observed in this bin among the subset of codes sharing the same prefix is incremented by 1;

(iii) 0 is attributed at each downstream bin from $p+1$ (inclusive).

Of note, when s_{ij} = 100%, the LIN code of the new profile *j* is given the complete LIN code of *i* (including at the last bin).

Missing data, equal matches and input order of profiles are handled as explained in **Box 2**.

The process of assigning a LIN code to a cgMLST profile first involves matching it against all existing defined LIN-encoded cgSTs to identify its closest neighbor (*i.e.*, the reference profile). If the two profiles (new and reference) have no dissimilarity (*i.e.*, no allele mismatch among the loci called in both profiles), the LIN code of the reference is simply assigned to the new profile. This will happen when the new cgST differs from the reference only by its missing data pattern (see **Box 2**). Otherwise, when the two profiles differ by at least one allele, a novel LIN code is created. For this, the pivot bin is defined as the bin in which the observed allele dissimilarity falls, and the novel LIN code is created in three steps (**Figure 1; Box 1**): (i) copying the LIN code prefix of the reference isolate, *i.e.* from the left bin up to the pivot bin (excluded); (ii) incrementing by 1 the maximum integer value observed in the pivot bin among the profile(s) sharing the same prefix used at step (i); (iii) attributing the integer value 0 at the bins downstream of the pivot, corresponding to initialization of the novel subdivision created at the pivot bin level.



	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 X	**Y** (5.00%)	0	2	0	0	0	0	0	0	0	0	
Genome Z	**X** (98.90%)	0	2	0	0	0	0	1	0	0	0	
	Coding steps (genome Z):	(i)						(ii)	(iii)			

Figure 1. **Overview of the process of cgMLST-based LIN code assignment.** The process starts with assigning cgMLST profiles to genome sequences and classification of profiles into unique core genome sequence types (cgST). After an initialization step (full-0 code for the first cgST), LIN codes are created for each cgST using the similarity to its closest-related already encoded cgST (steps i, ii and iii; see details in main text and Box 1). The bins and their threshold values are those chosen for the KpSC. The asterisk (*) indicates that the values are for the right threshold of each bin, exclusive. Note that there is no bin corresponding to complete similarity (gray column on the right), as in this case the LIN codes are identical, *i.e.*, there is no need to create a novel LIN code.

A LIN code prefix can be defined as any bin subset that starts from the leftmost position of the complete LIN code. The notion of prefix is important as it conveys a sense of genetic similarity among profiles: the longer the common prefix of two LIN codes is, the more similar the two corresponding profiles are. For a given cgST profile, its LIN code thus expresses how similar it is to other cgMLST profiles. Very different profiles will show identity at few or no prefix positions of their LIN codes, whereas nearly identical genomes will have LIN codes identical at most or all positions (see *e.g.*, **Figure 1**, genomes Z *versus* X: shared prefix 0_2_0_0_0_0 implies a minimum similarity of 98.88%, inclusive, and a maximum similarity of 99.36%, exclusive). Of note, our definition of LIN code prefix is similar to the LINgroup concept proposed by Vinatzer and colleagues (Vinatzer et al., 2017).

An important particularity of LIN codes is that the numerical identifiers at a given bin position (except the leftmost one) can only be interpreted in the context of the LIN code prefix preceding the considered bin: the same integer value at a given bin position corresponds to group membership only if the upstream prefixes are identical. In other words, groups at a given bin position are subdivisions of the upstream prefixes and are numbered starting from zero independently for each prefix. This particularity of LIN codes reduces the total number of integer identifiers observed in each position, making them easier to read than systems in which a group identifier is created independently at each level (for example, there are currently > 10,000 group identifiers at HierCC-1 level; (Achtman et al., 2022)). Interestingly, the diversity observed within a group defined by a given prefix can immediately be deduced from the maximal integer found among its members in the bin immediately downstream of the prefix length (**Figure 2**).

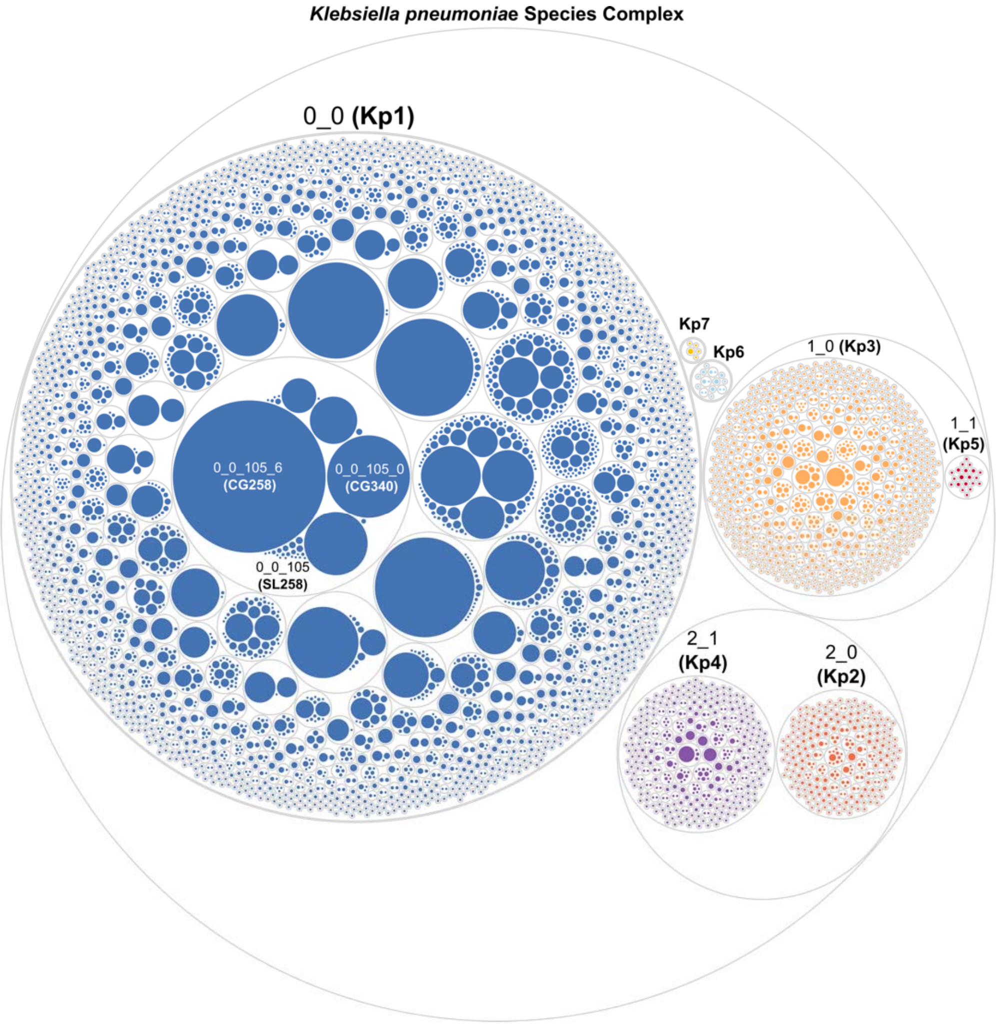


Figure 2. **The hierarchical nature of LIN code positions**. Numbering starts from 0 for subdividing each higher-level partition, characterized by a unique LIN code prefix. The hierarchical structure of LIN codes is shown here with a circular packing plot obtained from the KpSC data from BIGSdb-Pasteur. The circles correspond to LIN code prefixes of lengths 1 to 4 (an extra, all-encompassing circle corresponds to the entire KpSC); the size of the circles is related to the number of genomes they comprise. The first two bins in the LIN code prefix are used to identify phylogroups. Where for some phylogroups the first bin is unique (*e.g.*, prefix 0 for Kp1), in other cases it is common to multiple phylogroups (*e.g.*, prefix 2, which is associated with both Kp2 and Kp4), and therefore the second bin is necessary to discriminate between them (*e.g.*, 2_0 and 2_1 for Kp2 and Kp4, respectively). The hierarchical nature of LIN codes applies to subsequent levels of the prefix such as to those corresponding to sublineages (third bin, e.g. Kp1 SL258 is identified with the LIN code prefix 0_0_105) and to clonal groups (fourth bin, e.g. Kp1 CG258 with the LIN code prefix 0_0_105_6). Data was plotted in R v4.3.2 with ggplot2 and edited using Inkscape.

Box 2. The particulars of LIN codes: handling of missing data, equal matches, input order and computational precision

Missing data. Whereas 7-gene MLST genotyping requires complete allelic profiles, cgMLST approaches can tolerate the presence of missing alleles, as some core genes may not be essential, and as genome assembly shortfalls occasionally result in the absence or incompleteness of some loci. Therefore, the definition of cgSTs needs to accommodate missing data. Profiles may differ only by loci where there is one or more missing allele(s) in one of the profiles, while otherwise identical at all loci called in both profiles. Such profiles will be assigned to distinct cgSTs. We define as coincident cgSTs, groups of cgST profiles that differ only by their missing data pattern. As the dissimilarity between profiles is computed based solely on loci called in both profiles (Hennart et al., 2022), coincident cgST profiles will have a 0 dissimilarity value between them, and therefore the same LIN code.

Near-identical isolates or different WGS runs of the same isolate can lead to variable missing allele calls but are otherwise identical in the called loci, and will as a consequence lead to the creation of two or more coincident cgSTs. Each of these isolates' profiles will match with these multiple coincident cgST. When a given profile matches two or more predefined coincident cgSTs, it will (by definition) be attributed to all the coincident cgSTs. To minimize this phenomenon, a maximum number of accepted missing data must be defined when implementing the cgST classification within BIGSdb.

Equal matches and unicity of LIN codes. As described above, an isolate's profile may match more than one encoded cgST, due to missing loci. In this case, a unique LIN code will be defined (and displayed) for the isolate. To choose between the different possibilities, the LIN code of the cgST with the fewest missing allele(s) will be attributed. When two or more coincident cgSTs have the same number of missing allele(s), the cgST with the smallest LIN code partition identifiers (considered from left to right bin, *i.e.*, the lowest sort order) will be chosen. The same priority rule is applied to encode every novel profile that is equidistant to two (or more) previously LIN-encoded non-coincidental cgSTs.

Input order. The LIN code approach is dependent on input order, as the partition in a given bin may vary slightly according to the order by which the genomes were encoded (Hennart et al., 2022). To minimize this effect, BIGSdb uses the traversal of a minimum spanning tree (MStree; (Prim, 1957)) to define the order by which the novel profiles are encoded. To code a novel batch of genomes, after creating a MStree, the isolate chosen as the starting point for LIN encoding is the one that has the closest similarity to an already encoded isolate in the database; next, the MStree is traversed from this node. This approach (implemented since v1.36.1) maximizes reproducibility when adding a batch of novel genomes. To minimize the number of resulting prefix-based partitions, novel genomes should be encoded in batches as large as possible.

Computational precision. As for all categorizations that rely on thresholds, computational precision is critical for reproducible results. For example, the pairwise dissimilarity between cgMLST profiles, which is a ratio, may often have a higher number of decimals than can be handled by the computing system, and its rounded value may lead to a slight underestimate (or overestimate) of the true value. When the (true) dissimilarity between an incoming profile and its reference is exactly identical to the left threshold of a bin (*i.e.*, the same ratio of distinct versus called alleles), a rounded value may incorrectly correspond to the previous bin (**Figure 3**). Therefore, pairwise dissimilarity computations should be performed in a way exactly identical to the bin thresholds themselves. In BIGSdb, ratios corresponding to the thresholds are compared to the calculated dissimilarity values using Perl platform-native floating point values (usually IEEE 754 double-precision).

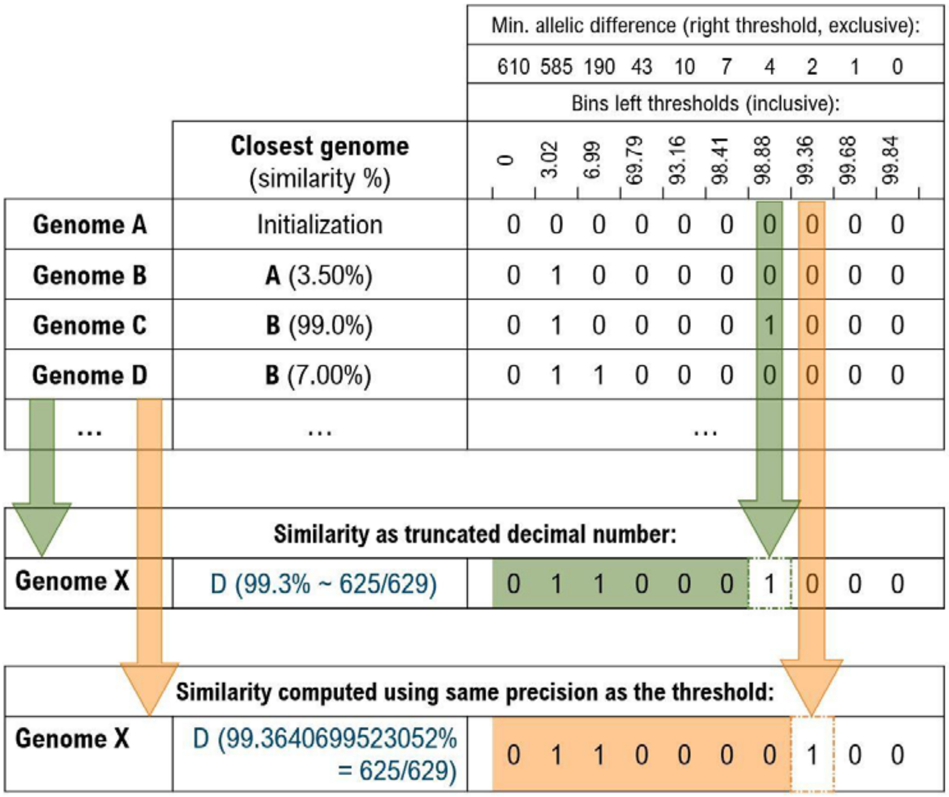

Min. allelic difference (right threshold, exclusive):
610 585 190 43 10 7 4 2 1 0
Bins left thresholds (inclusive):
0 3.02 6.99 69.79 93.16 98.41 98.88 99.36 99.68 99.84
Closest genome (similarity %)
Genome A | Initialization | 0 0 0 0 0 0 0 0 0 0
Genome B | A (3.50%) | 0 1 0 0 0 0 0 0 0 0
Genome C | 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
… | … | …
Similarity as truncated decimal number:
Genome X | D (99.3% ~ 625/629) | 0 1 1 0 0 0 1 0 0 0
Similarity computed using same precision as the threshold:
Genome X | D (99.3640699523052% = 625/629) | 0 1 1 0 0 0 0 1 0 0

Figure 3. **The effect of rounded cgMLST similarity values on LIN code assignment**. In this example, the use of a rounded value for the similarity between genome X and genome D leads to a slight underestimate, therefore creating a novel identifier in bin 7, instead of bin 8 when computing the similarity with the same precision as the threshold.

LIN codes functionalities implemented within the BIGSdb platform

The LIN code taxonomy of KpSC genomes was incorporated into the Institut Pasteur *K. pneumoniae* MLST and whole-genome MLST platform (https://bigsdb.pasteur.fr/klebsiella), using BIGSdb v1.34.0 and upwards (Hennart et al., 2022). For the KpSC, this database plays the role of the source database for the definitions of alleles, cgMLST profiles, cgSTs, and LIN codes.

In BIGSdb, LIN code schemes can be defined in the curator's interface of both the 'sequence definition' and 'isolates' databases. A LIN code taxonomy is created with reference to a defined indexed scheme, *e.g.*, cgMLST. An indexed scheme is a scheme with a unique identifier for each profile, *e.g.*, cgST here. To index a scheme, one needs to specify the maximum number of missing alleles accepted for profiles to be assigned to cgSTs. To create a LIN code taxonomy, allele mismatch thresholds that define the LIN code bins must simply be defined. In the case of KpSC, the 629-loci cgMLST scheme was selected, and ten thresholds were defined (**Figure 1**).

Users who wish to assign a novel LIN code for a KpSC isolate must submit the genome sequence(s) to the BIGSdb-Pasteur 'isolates and genomes' database. If all quality criteria are fulfilled (https://bigsdb.pasteur.fr/klebsiella/genome-quality-check/), the genome(s) will be deposited in the database for allele, cgMLST profile, cgST and LIN code definitions. The inferred cgMLST profiles, as well as their cgST identifiers and LIN codes, will be made openly accessible through the sequence and profile definition database ('seqdef'). To ensure confidentiality of users' data when requested, isolate metadata and associated genome sequence(s) can be embargoed and released at a later stage.

Users can search *K. pneumoniae* isolates of interest using the LIN code matching functionalities implemented in BIGSdb. A complete LIN code (or any prefix) can be used as a query. The nickname nomenclature attached to LIN code prefixes can also be used to facilitate the query of groups of interest (*e.g.*, SL258 members can be searched by using its attached prefix 0_0_105, or using the SL258 nickname itself). The list of genomes from the query results can be further analyzed using the available analytical tools within the BIGSdb platform, or exported for external use.

Section 2: Novel developments and examples of applications

Multiple *Klebsiella pneumoniae* 7-gene MLST sequence types are polyphyletic

Even though they are based on allelic profile comparisons rather than a sequence-based phylogenetic analysis, LIN code prefixes of length 3 or 4 bins are compatible with phylogenetic classifications and thus represent markers of their corresponding tree branches (Hennart et al., 2022). In contrast, 7-gene MLST may conflate phylogenetically unrelated genomes in a single ST, for example through recombination leading to the same ST being assigned to genomes from distinct parental lineages, or by large recombinations affecting multiple cgMLST loci but leaving the 7-gene MLST loci unaffected (Lam et al., 2023). Here we explore the extent of this phenomenon using 44,000 publicly available genomes of *K. pneumoniae* (June 2023). We found that 113 STs are polyphyletic, defined here as being observed in at least two unrelated LIN code sublineages (**Table S1**). We illustrate this phenomenon for major STs in **Figure 4**. For example, ST485 was observed in four phylogenetically unrelated sublineages: SL485 (0_0_157), SL45 (0_0_158), SL1626 (0_0_227) and SL11569 (0_0_1215). ST347 stands out as being observed in 8 distinct sublineages. This analysis also confirmed the polyphyletic status of ST23 (Lam et al., 2023), which conflates isolates from distant sublineages: SL23 (0_0_429) and SL218 (0_0_115).

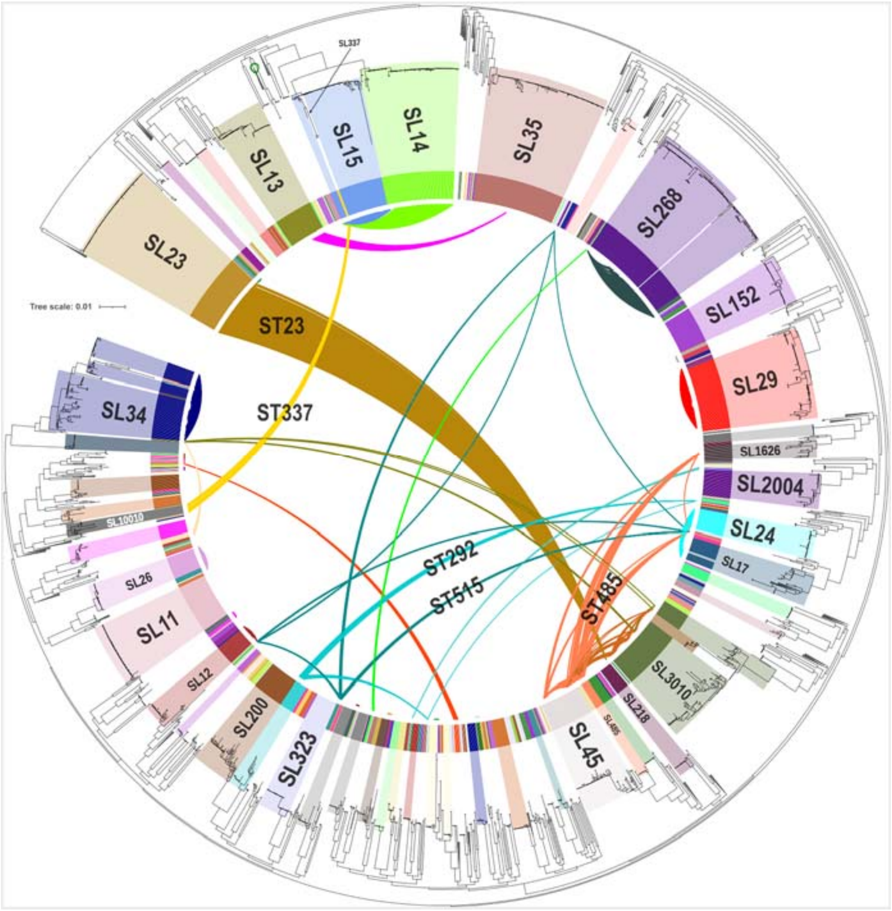
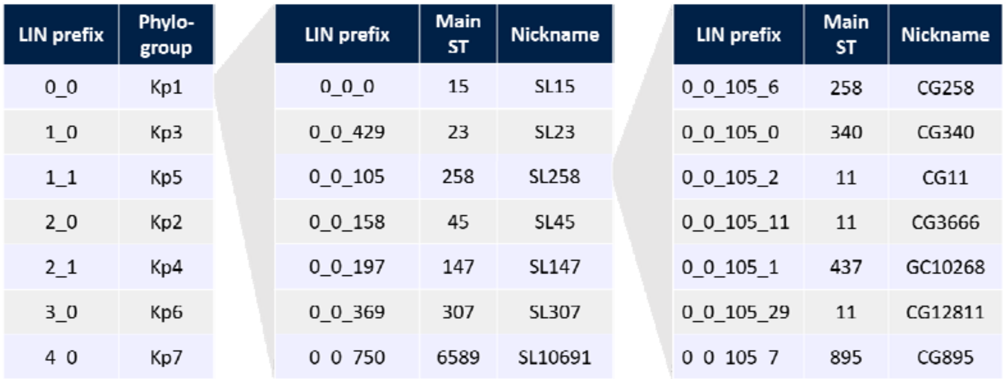


Figure 4. Phylogenetic tree of *K. pneumoniae* main sublineages. A phylogenetic analysis of 5,665 *K. pneumoniae sensu stricto* genomes (LIN code prefix 0_0; see selection process in Methods) was performed from the multiple sequence alignments of 629 cgMLST genes. Closely related leaves were collapsed. The colored sectors in the inner circle correspond to the sublineages (SL) defined based on their prefix of length 3 (*i.e.*, made of the three first bins); the major sublineages are highlighted by lighter-colored sectors joining the circle to the tree leaves. The internal connectors between sublineages represent frequent STs that were found in two or more sublineages. The full interactive tree is available at: https://itol.embl.de/tree/1579917420525181688029926

Nicknaming the LIN code prefixes enables carry-over of MLST identifiers into the genomic taxonomy

Whereas LIN code prefixes themselves can be used as canonical markers of groups of interest that are easy to handle by computers, for humans, prefixes are not very easy to remember or pronounce. Here, we propose to nickname the LIN code prefixes with simple denominations using a LIN code prefix nicknaming system (newly implemented within BIGSdb; https://bigsdb.readthedocs.io/en/latest/administration.html?highlight=prefix#setting-up-lincode-definitions-for-cgmlst-schemes). It is thereby possible to nickname every prefix in any chosen way, for example by incrementing an integer identifier for each novel prefix of a given length, analogous to the numbering of 7-gene MLST STs. Other labels could be applied, such as Greek letters, astronomical objects, or any other series of words that may be universally understandable and easy to remember. This nicknaming process would be particularly useful for long prefixes, or prefixes of particular relevance that subdivide the population at particularly informative levels.

For bacterial species where previous nomenclatures exist, a novel and unrelated naming system would have the drawback of creating yet another nomenclature. Assigning nicknames to prefixes based on the previous nomenclature system is therefore more meaningful. For *K. pneumoniae*, the classical MLST nomenclature system is widely used, and knowledge has accumulated on the epidemiological history and characteristics of predominant STs. We therefore aimed to create backward nomenclatural compatibility of LIN codes with ST identifiers. We used a majority identifier inheritance rule that was previously developed and applied to single-linkage cgMLST groups (Hennart et al., 2022). We applied this approach to nickname LIN code prefixes of lengths 3 and 4 bins (which, for convenience, we have defined as sublineages and clonal groups, respectively) by using ST identifiers as a source. In short, for each LIN code prefix of length 3 or 4 bins, the identifier of the predominant ST among its genomes was used as a label, wherever possible (*i.e.*, if not yet attributed). Following this approach, most SLs and CGs were indeed labeled according to the ST identifiers of most of their isolates, whereas a minority are nicknamed with incremental numbers (because the majority ST was already used for another prefix). In **Figure 5,** we provide illustrative examples of correspondence between prefixes and nicknames for major clonal groups. For example, ST258, and its derivative ST512, share the prefix 0_0_105, nicknamed SL258, and the 4-positions prefix 0_0_105_6, nicknamed CG258.



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

Figure 5. Nicknaming of LIN code prefixes enables inheritance of previous nomenclatures. Nicknames of some LIN code prefixes of lengths 2 to 4 bins, inherited from phylogroup numbering or Linnaean taxonomy (2-bin prefix, left panel) or 7-gene MLST (prefixes of lengths 3 and 4 bins, central and right panels), are displayed.

Note that the MLST nickname inheritance rule was applied only using ST identifiers up to ST6500 (**Figure S1**). Given that the main sublineages of KpSC have long been sampled, the inheritance of MLST identifiers on SL and CG identifiers will apply to most of the extant diversity of the KpSC. For subsequent prefixes, SL and CG nicknames are numbered incrementally, starting with 10,000 (see example on **Figure 5**) in order to make clear that these new nicknames are not inherited from MLST nomenclature. In parallel, continual expansion of the MLST nomenclature will result in defining STs (incremented by one) upwards of 6500 (currently the highest ST is ST6859, January 21st, 2024). Hence, a correspondence between ST identifiers >6500 and prefix nicknames >10,000 may exist but will not be immediately obvious. For novel sublineages and clonal groups that may emerge in the future, our recommendation is to prioritize their LIN code SL and CG nicknames, rather than their ST, when communicating on these groups. Note that the 2-bin prefixes of *Klebsiella* LIN codes each define a particular KpSC phylogroup, corresponding to the seven currently described species or subspecies (Hennart et al., 2022), and were thus nicknamed accordingly (**Figure 5**).

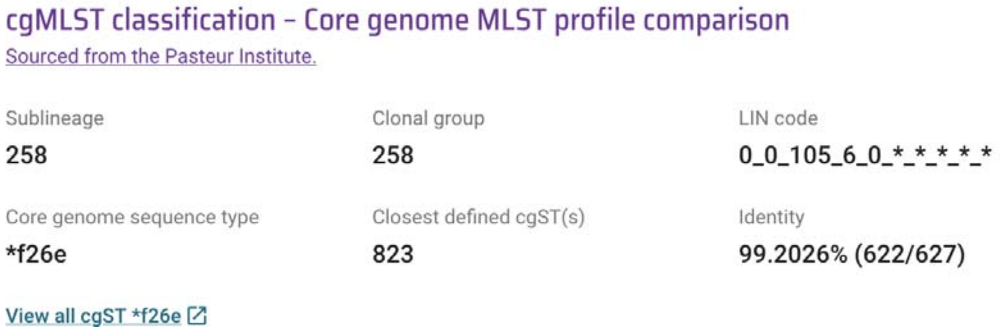
From dual- to single-barcoding taxonomy of *Klebsiella pneumoniae* strains

Previously, cgMLST groups were defined by the single-linkage (slink) clustering method using the same 10 thresholds as for the LIN codes, and the four highest-level groups were nicknamed by inheritance from Linnaean taxon names (for the two first) or MLST labels (for the levels defined by thresholds 190 and 43, dubbed Sublineage and Clonal Groups, respectively) (Hennart et al., 2022). Together with the LIN code taxonomy (which had no nickname in (Hennart et al., 2022)), this slink-based system formed a 'dual-barcoding approach'. However, because such slink groups suffered from fusion of existing groups upon addition of subsequent genotypes, which occasionally had intermediate distances between preexisting groups (*e.g*., hybrid genotypes), the classification of cgMLST profiles into slink groups was abandoned. Fortunately, when excluding the hybrid genotypes, a nearly complete concordance was observed at the four first levels between slink clusters and LIN code groupings optimized based on MStree (Hennart et al., 2022). As a result, the LIN code taxonomy currently in use is nearly fully consistent with the one initially proposed (only SL10000 to SL10021, and CG10000 to CG10276 correspond to groups that were renamed; table of correspondence available upon request). The use of a single-barcoding taxonomic system based on LIN codes will stabilize and simplify the way groups are defined and labeled.

LIN code taxonomy usage in external genomic epidemiology platforms

To make the LIN code taxonomy accessible for external tools, databases and analysis platforms, the LIN code nomenclature components (alleles, profiles, cgSTs and LIN codes) can be extracted from BIGSdb using an application programming interface (Jolley et al., 2017). This can be performed via a single query using the following link: https://bigsdb.pasteur.fr/api/db/pubmlst_klebsiella_seqdef/schemes/18/profiles_csv. However it is important to note that to be effective, external copies of the database need to be very frequently synchronized with the primary nomenclature database. This is because, when genome sequences (through their cgMLST profiles) are matched to the LIN code taxonomy, an incomplete LIN code may be defined in many cases, as no identical cgMLST profile may be existing at this time in the source LIN code taxonomy. In such cases, a new nomenclatural identifier must be defined and assigned, but this is only possible within the source database otherwise consistency of nomenclature will be lost. Inference of the query genome's LIN code in external resources can only be inferred up to the bin preceding the pivot bin corresponding to the closest match. Notably though, when the LIN code prefix up to the fourth bin (at least) can be defined for the query genome, information on species, subspecies, SL and CG can be derived. If the query genome is closely related to one in the source database, its LIN code will be almost completely defined. Therefore, although novel cgMLST alleles, cgST profiles and LIN codes can only be defined in the source database of the nomenclature (BIGSdb-Pasteur for the KpSC), the use of LIN codes in external databases or tools still has functional relevance. For any genome (cgMLST profile) that has no complete LIN code, data submission to the source database is encouraged, in order to update the LIN code taxonomy and define complete LIN codes for the novel genomes.

To illustrate the external use of LIN codes, we implemented the KpSC LIN code taxonomy in the Pathogenwatch platform, in which a KpSC database was set-up previously (Argimón et al., 2021). First, on a regular basis, Pathogenwatch synchronizes from BIGSdb into its internal temporary database, the defined alleles, cgSTs and associated LIN codes, using the API functionality of BIGSdb. Second, the cgMLST allele sequences extracted from the query genome assembly are compared to those in the temporary database, and the cgMLST profile is used to find the closest match in the temporary database. If the query genome does not match completely with an existing source nomenclature cgST, a provisional cgST is assigned, represented by the asterisk and a code (*e.g.*, cgST *f26e). Pathogenwatch also indicates the closest cgST defined in the source taxonomy database and provides a link to the list of all isolates within Pathogenwatch that have the same cgST genotype. Third, an incomplete LIN code will be provided by Pathogenwatch based on the shared prefix with the closest reference cgST (**Figure 6**). This process provides information about the relatedness of a query Pathogenwatch genome compared to the existing taxonomy elements and can in most cases provide sublineage and clonal group identification. In those cases where Pathogenwatch provides provisional alleles, STs, cgSTs and/or LIN codes, the user is encouraged to submit the genomic sequence data to the source BIGSdb-Pasteur database so that novel nomenclatural identifiers (alleles, STs, cgSTs, LIN codes) can be created. Note that as Pathogenwatch uses its own algorithm to provide the species and subspecies for KpSC genomes, this taxonomic information is not deduced from LIN codes in that platform.



cgMLST classification – Core genome MLST profile comparison

Sourced from the Pasteur Institute.

Sublineage	Clonal group	LIN code
258	258	0_0_105_6_0_*_*_*_*_*
Core genome sequence type	Closest defined cgST(s)	Identity
*f26e	823	99.2026% (622/627)

View all cgST *f26e

Figure 6. Example of LIN code identification in Pathogenwatch. Although the LIN code is incomplete, the genome can be inferred to belong to clonal group 258 (defined as prefix 0_0_105_6), which comprises ST258 and ST512 isolates (see **Figure 7**).

Applications of LIN codes to subdivisions within high-risk Kp sublineages

A number of *K. pneumoniae* sublineages, including SL258, SL147, SL307, SL17 and SL23, have been recognized to cause a large burden of so-called hypervirulent or multidrug resistant infections. These groups have been the subject of detailed studies, that have led to defining their geographical spread and phylogenetic subgroups (Deleo et al., 2014; Hetland et al., 2023; Lam et al., 2018; Rodrigues et al., 2022; Wyres et al., 2019). However, so far, a harmonized nomenclature of these subgroups has been lacking, making it difficult to recognize them in subsequent studies. Here, we illustrate how LIN codes can help track Kp dissemination at fine genetic scales within sublineages, using the example of SL258, a major *Klebsiella pneumoniae* carbapenemase (KPC) producing sublineage of *K. pneumoniae*.

SL258 is defined by its LIN code prefix, 0_0_105, and encompasses all isolates from 7-gene ST11, ST258, ST340, ST512 and some others (**Figure 5**). Its phylogenetic structure shows that SL258 is divided into several clades (**Figure 7**) that are labeled with their unique clonal group number. These include CG258 (0_0_105_**6**), defined by LIN code position 4, which contains all ST258 and ST512 isolates. LIN code position 5 can further be used to distinguish major subclades within SL258, including ST340 (0_0_105_**0_11**) and ST437 (0_0_105_**1_1**) and other subclades within ST11, some of which appear to be associated with recombination events that include the capsule locus (KL column in **Figure 7**). The LIN codes can also help distinguish between different subclades that are associated with the same capsule locus. For example, they clearly distinguish 3 subclades that are all ST11-KL64 (grey shading on the tree branches, **Figure 7**). One of these is the major lineage circulating in China (0_0_105_**2_0_0_2**, predominantly 0_0_105_2_0_0_2_17, 24/30 genomes) that carries KPC-2 and often the *iuc1* aerobactin virulence locus, descended from ST11-KL47-KPC-2 (0_0_105_**2_0_0_2_*,** where * is not 17), as discussed broadly in the literature (Zhou et al., 2023, 2020). A second, unrelated ST11 subclade carrying KL64 (0_0_105_**0_0**) is circulating in South America (encoding KPC-2, but rarely *iuc*), while a third smaller clade (0_0_105_**0_2**) is detected primarily in Taiwan rather than in mainland China (lacking KPC and with only one of eight genomes carrying *iuc*). The example of SL258 illustrates how LIN code classification beneath the sublineage level can help recognize and name subgroups of medical and epidemiological relevance, which should be the object of enhanced surveillance.

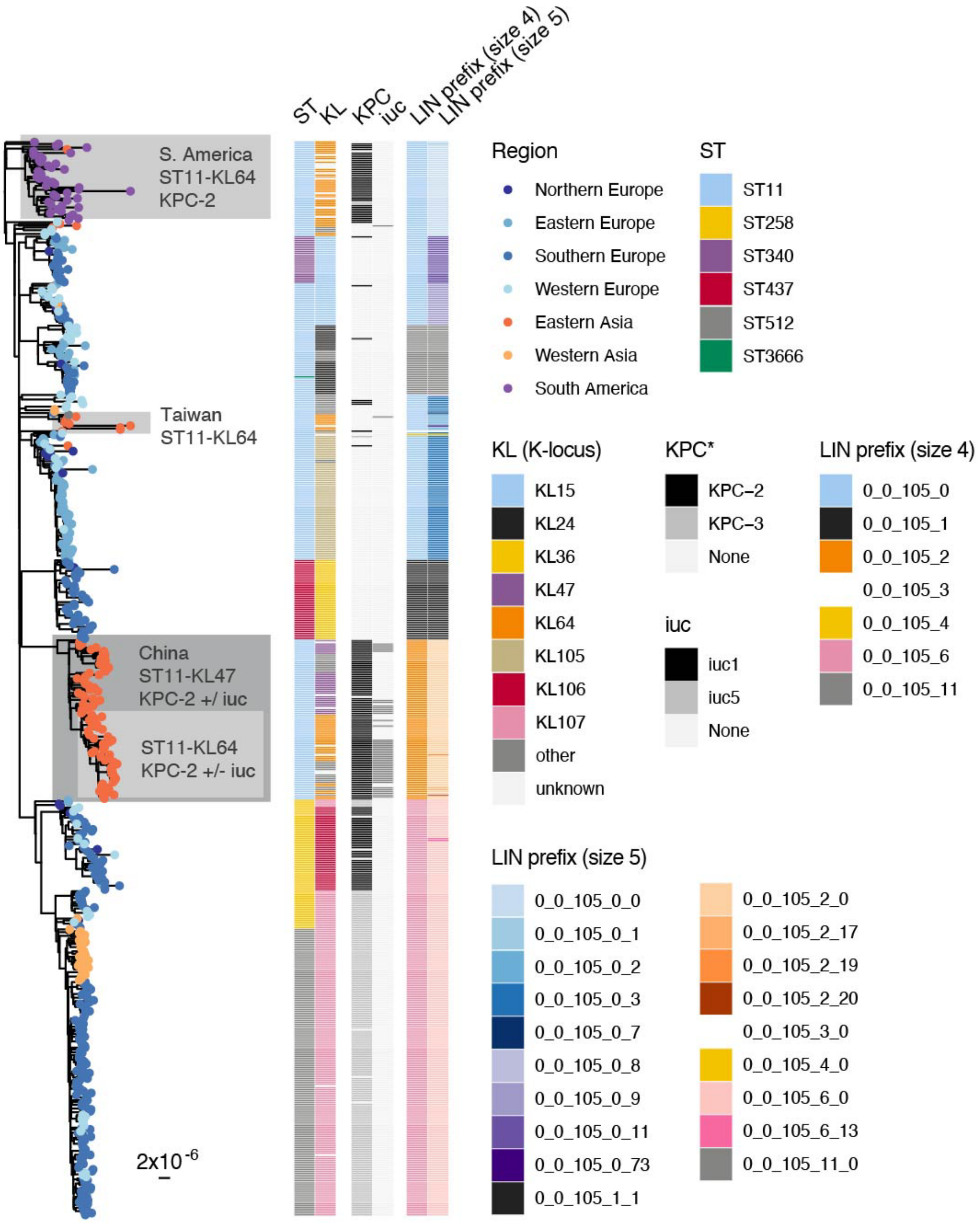


Figure 7. SL258 phylogenetic structure and LIN codes. Maximum-likelihood phylogenetic tree of n=586 SL258 genomes inferred from a recombination-free variable site alignment (see Methods). Tips are coloured to indicate geographic region of origin as per the legend (United Nations region classifications). The distribution of 7-gene sequence types (STs), K-loci (KL), *bla*$_{\mathrm{KPC}}$ (KPC) alleles, aerobactin locus lineages (*iuc*), LIN code prefixes of sizes 4 and 5, are indicated by colored blocks as per the legends (note that colors are independent to each column). Only K-loci identified with a Kaptive confidence score of 'Good' or better are shown (otherwise marked 'unknown'). Two isolates were detected with *bla*$_{\mathrm{KPC-30}}$ and one with *bla*$_{\mathrm{KPC-12}}$ but are not shown in the figure for brevity. Sub-clades described in the text are coloured and labeled accordingly.

Application of LIN codes to outbreak strain identification

To illustrate the use of LIN codes to identify outbreak strains, and to track strain diversification during protracted outbreaks, we explored the example of SL147. This is a prominent multidrug-resistant international sublineage of *K. pneumoniae*, defined by its LIN code prefix 0_0_197. **Figure S2** illustrates how the phylogenetic relationships within SL147 are captured by LIN codes, using a previously described dataset (Rodrigues et al., 2022). SL147 comprises a single clonal group (0_0_197_0) and three 7-gene STs (ST147, ST273 and ST392). At LIN code position 5, four partitions (0_0_197_0_**0**, 0_0_197_0_**4**, 0_0_197_0_**17** and 0_0_197_0_**25**) correspond largely to ST273, ST392 and two deep branches of ST147. In addition, both ST147 and ST273 are genetically heterogeneous and structured phylogenetically into several minor branches, which were captured by additional partitions of LIN code level 5 (**Figure S2, panel A**).

Protracted outbreaks often lead their investigators to define local clades (or subgroups) within the closely related outbreak isolates. These clades are often attributed temporary placeholder names, which are difficult to compare across studies *e.g.*, Clade A and Clade B (Martin et al., 2021). We illustrate how LIN codes provide a way to define these clades definitively, using the diversity among outbreak isolates from a metallo-β-lactamase (NDM)–producing carbapenem-resistant ST147 outbreak in Tuscany (**Figure S2, panel B; Table S2**). The time span of the Tuscany outbreak is November 2018 **-** 2021. Most of the isolates in this outbreak have prefix 0_0_197_0_**4_1_0**, thus differing by no more than 4 alleles out of 629 with another member of the group. The authors defined two clades, A and B. Here, clade B corresponds to the set of LIN codes 0_0_197_0_4_1_0_8_x_x (*i.e.*, with prefix 0_0_197_0_4_1_0_**8**, with x meaning there may be variation at the two last positions). Clade A was more diverse, and LIN codes classify this genetic variability in a definitive way, with six 8th position prefixes (0_0_197_0_4_1_0_**7**, 0_0_197_0_4_1_0_**9**, 0_0_197_0_4_1_0_**10**, 0_0_197_0_4_1_0_**11**, 0_0_197_0_4_1_0_**12** and 0_0_197_0_4_1_0_**66**). This example highlights how *K. pneumoniae* LIN codes can subdivide isolates from long-term outbreaks.

A search of the BIGSdb-Pasteur KpSC database (January 31st, 2024) for prefix 0_0_197_0_4_1_**0** identified n=395 *K. pneumoniae* genomes, isolated between 2014 and 2023 and coming from 20 countries from North America, Europe, Asia, Africa and Oceania, which indicate the global dissemination of this particular subgroup of SL147. However, prefix 0_0_197_0_4_1_**0_8** was so far only reported from the Italian outbreak. This example illustrates how LIN codes can facilitate the tracking of strain dissemination, by enabling the identification of similar isolates from separate studies. As an outbreak strain prefix can be easily discussed and shared among investigators and is sufficient to exchange information on strain identity across countries, LIN codes enable genomic surveillance investigations without the need to share genomic sequences, which may alleviate issues around data confidentiality. Likewise, for the surveillance of particularly concerning strains, early warnings could be triggered based on the detection of the specific LIN code of the strains under surveillance.

Given that LIN codes are phylogenetically informative, they can be represented graphically as prefix trees, which broadly approximate the phylogenetic relationships among isolates (Hennart et al., 2022). Here, we introduce the tool LINtree to create prefix trees from LIN codes (https://gitlab.pasteur.fr/GIPhy/LINtree). The input file contains a list of genome names and LIN codes (one sample per row), with a header row indicating the level of similarity for each bin. LINtree outputs a Newick-formatted tree showing the relationships between input genomes, based on the hierarchy provided by the LIN codes and with branch lengths scaled using the similarity levels in the header row. For example, the tree of the ST147 Italian outbreak shown in **Figure S2** was generated using this tool, based on the input list of LIN codes. This example illustrates how the prefix tree recapitulates the phylogenetic relationships of this outbreak strain with its ancestral relatives, providing a useful aid in outbreak investigations.

Discussion

Facilitating communication on the intraspecific diversity of bacterial strains is a key objective of strain taxonomies, which entail classification and naming of groups within species. In the field of epidemiological surveillance of pathogens, it has long been recognized that strain typing methods used for long-term and global strain tracking should rely on an internationally standardized nomenclature (Struelens, 1998). In turn, a robust and fine-grained strain taxonomy promotes the understanding of the links between genotypes and clinical phenotypes, vaccine coverage and antimicrobial resistance (Achtman et al., 2022; Maiden et al., 2013).

Here we have presented in detail the cgMLST-based LIN code approach and further developed this novel strain taxonomy system. The stability of LIN code classification is a critical property, which has been impossible to achieve with previous strain classification systems relying on single-linkage clustering (such as MLST clonal complexes defined by BURST or cgMLST single-linkage groups). cgMLST LIN codes are stable, as the incorporation of novel genomes has no effect on pre-existing LIN codes (**Figure 1**). Here, we have presented important enhancements of our initial implementation, by (i) improving the reproducibility of LIN encoding by addressing the dependency of this approach to rounded genetic distance values; (ii) the implementation within the BIGSdb platform, of input order rules for creating novel LIN codes, and (iii) implementing formal rules for handling missing data. These improvements optimize the definition of LIN codes and have resulted in a robust strain taxonomy system that is now in operation for *K. pneumoniae* since January 2023 and currently comprises 37,070 cgSTs and 32,500 LIN codes, which correspond to 2,492 sublineages and 4,230 clonal groups (January 28th, 2024).

In this work, we also extend the LIN code approach by proposing and implementing a nicknaming system for LIN code prefixes. As shown previously (Hennart et al., 2022; Marakeby et al., 2014), LIN codes are highly compatible with phylogenetic relationships, and their prefixes can therefore act as markers of phylogenetic groups. Nicknaming was designed to be flexible, and can thus accommodate any naming system of choice, either numerical or textual. To ensure continuity with 7-gene MLST nomenclature, we had previously proposed to nickname cgMLST single-linkage groups (Hennart et al., 2022). For *K. pneumoniae*, we had nicknamed the partitions within two special levels with thresholds of 43 and 190 mismatches, defined as “sublineages” and “clonal groups”, respectively. However, because of the instability of the single-linkage clustering approach, we soon observed fusions of previously defined (and nicknamed) groups, rendering the single-linkage-based nomenclature unstable. Here, we instead nickname the LIN code prefixes of lengths 3 and 4 bins, which correspond to the same thresholds as previously defined “sublineages” and “clonal groups”, respectively. Hence, we here redefined the “sublineages” and “clonal groups” as being based on LIN code prefixes.

A key property of a novel nomenclature system is its continuity with previous nomenclatures, as it minimizes confusion and facilitates its adoption by microbiologists and epidemiologists. Establishing a dictionary of correspondence between novel and previous nomenclatures is a possibility but it implies cumbersome handling of both series of identifiers. Here, we provide the possibility of embedding any previous nomenclature(s) within the LIN code taxonomy. In the case of *K. pneumoniae*, by using a previously described inheritance algorithm (Hennart et al., 2022) that has mapped the 7-gene ST identifiers onto LIN code prefixes of lengths 3 and 4 bins, we provide continuity between the novel nomenclature of sublineages and clonal groups with the widely used MLST standard. Using LIN code prefix nicknames instead of MLST identifiers has the additional benefit of enhancing the compatibility of the nomenclature with phylogenetic relationships: we have shown here for *K. pneumoniae* that classical MLST profiles often conflate unrelated sublineages. Note that we still recommend the maintenance and extension of the MLST nomenclature to classify future *K. pneumoniae* isolates, in parallel to the novel genomic nomenclature. However, we suggest the prioritization of LIN code nomenclature over MLST, which will be particularly important for sublineage and clonal group designations above 10,000 that are not inherited from MLST.

Hierarchical clustering (HierCC) also provides stable classifications and is likewise implemented based on cgMLST schemes (Zhou et al., 2021). Unlike for LIN codes, HierCC partition identifiers are incremented independently across levels, necessitating the handling of large integers, particularly in bins corresponding to the highest similarities, where over 100,000 partitions might be created. In contrast, LIN codes re-initiate the numbering from 0 within a bin, for each subdivision of a partition in the upper bin, resulting in a predominance of small integers, which are easier to handle for humans. By design, HierCC is stable only in its production mode, whereas it relies on the unstable single-linkage clustering approach in its development mode, implying an arbitrary decision on the switch from development to production to achieve stability.

LIN codes, as well as HierCC, are multilevel classifications that provide proxies of strain relationships. By conveying for each genome, its group membership and approximate degree of relatedness at various phylogenetic depths simultaneously, they are phylogenetically informative. LIN code prefixes are shared by genomes having at least the identity corresponding to the upper threshold of the last prefix bin (exclusive). The LIN codes (or HierCC codes) can in fact themselves be represented as a tree (formally, a prefix tree), with multifurcations corresponding to subdivisions of each prefix (**Figure S2, panel C;** see also (Hennart et al., 2022)) and node height corresponding to bin thresholds. This tree representation of LIN codes may serve as a proxy for the phylogenetic tree and can be created with no need of initial sequences or cgMLST profiles.

A taxonomic system needs to be created and updated in a coordinated manner. For this purpose, the cgMLST LIN code strain taxonomy approach was implemented in the BIGSdb platform. Its integration in this widely used platform will make it publicly available, and will facilitate its implementation for other bacterial species, as was recently illustrated for *Streptococcus pneumoniae* (Brueggemann *et al.* bioRxiv 2023, doi: https://doi.org/10.1101/2023.12.19.571883). The applicability to other bacterial species should be straightforward, provided that they comprise meaningful cgMLST diversity, excluding the so-called monomorphic pathogens (Achtman, 2008), such as *Mycobacterium tuberculosis* or *Salmonella enterica* serotype Typhi. Setting up LIN codes for other species will require defining tailored bin thresholds based on population structures, which requires globally representative genome datasets (**Figure S3, overview chart**). The approach could also be extended with minor adaptations to other organisms with predominantly clonal reproduction, such as protozoan parasites and fungi, even if they are not haploid (Bougnoux et al., 2004; Yeo et al., 2011). The wide adoption of the standardized cgMLST LIN code strain taxonomy would result in a universal strain nomenclature approach that could greatly enhance microbial biodiversity studies, international genomic epidemiology and infectious disease surveillance.

Methods

Identification of MLST sequence types that are discordant with sublineage classification

We used the 44,000 public genomes available in BIGSdb in June 2023. To spot potential discordances between ST and LIN code prefixes, we first filtered out non-Kp1 phylogroup (prefix 0_0) genomes and removed nearly identical cgMLST profiles, by keeping a single representative of each partition at LIN code bin 5. STs observed only in a single isolate were filtered out. We next searched for all STs that were split in several clonal groups or sublineages (as defined by their prefix) and conversely, also looked for prefixes of length 3 or 4 which comprised several STs. We then placed these genomes in a phylogenetic tree built using IQtree v2.2.2.2 (Minh et al., 2020) using GTR+I+G model, from concatenated alignments of individual cgMLST gene alignments.

SL258 phylogeny

Whole genome sequences representing SL258 were identified among the EuSCAPE collection (David et al., 2019) and two recent studies reporting 7-gene ST11 with K-locus (KL) 47 and KL64, for which multiple independent evolutions have been reported (Wang et al., 2023; Zhou et al., 2023). The ST11 genomes were subsampled to a manageable number as follows: (i) five randomly selected genomes per study year for each of ST11-KL47 and ST11-KL64 reported from sites across China, plus all ST11 with other K-loci reported in the same study included for context, total 92 genomes from this study (Zhou et al., 2023); (ii) 64 genomes representing ST11-KL64 clade 1 as defined in an analysis of public ST11-KL64 genomes (Wang et al., 2023). Genome assemblies were acquired from Pathogenwatch and those with >500 contigs and/or total assembly size < 4,969,898 or > 6,132,846 bp were removed (as per the KlebNET-GSP quality control definitions). Kleborate v2.3.2 (Lam et al., 2021) was used to determine 7-gene ST, *bla*$_{\text{KPC}}$ alleles, and *iuc* lineages (aerobactin locus), and Kaptive v2.0.7 (Lam et al., 2022; Wyres et al., 2016) was used to identify KLs.

In order to infer a high-resolution phylogenetic tree, genome assemblies were used to simulate 100bp paired end reads with wgsim (without errors, https://github.com/lh3/wgsim). Reads were mapped against the NJST258-1 completed reference genome (NCBI accession: CP006923.1) and single nucleotide variants called using the RedDog pipeline (https://github.com/katholt/RedDog). The resultant allele table was converted to a pseudo-whole genome alignment and used as input for Gubbins v2.3.2 (Croucher et al., 2015), in order to detect and remove recombination (100 iterations). The final filtered alignment of 10,390 variable sites, representing 591 genomes, was used to infer a maximum likelihood (ML) phylogenetic tree using RAxML v8.2.9 with parameters: best of 5 runs, 1,000 bootstraps each, gamma model of rate variation (Stamatakis, 2014). Subsequently, five genomes were removed due to excessive branch lengths. The ML tree was visualized with R v4.3.1 and the following packages: ape v5.7.1 (Paradis et al., 2004), phytools v 1.9-16 (Revell, 2024), and ggtree v 3.8.2 (Yu et al., 2017).

Data availability

There are currently 39,506 genomic sequences publicly available in BIGSdb. Genomes can be downloaded from the sequence bin page, and LIN codes are available in the main table retrieved following an isolates search page results. The complete and up-to-date LIN code nomenclature (comprising alleles, profiles, cgSTs and LIN codes) can be extracted from BIGSdb using a single query at the following link: https://bigsdb.pasteur.fr/api/db/pubmlst_klebsiella_seqdef/schemes/18/profiles_csv.

Acknowledgements

We thank François Lebreton for providing the Newick file of the Italian ST147 outbreak described in Martin *et al*., 2021. We acknowledge the help of the HPC Core Facility of the Institut Pasteur for this work.

Funding

This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation [INV-025280], Institut Pasteur and by European Union's Horizon 2020 research and innovation programme. This work was also supported financially by the French Government's Investissement d'Avenir program Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10-LABX-62-IBEID). BIGSdb development is funded by a Wellcome Trust Biomedical Resource grant (218205/Z/19/Z).

Under the grant conditions of the funders, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Authors contributions

Klebsiella genomic platform integration (KlebNET) conceptualization and coordination: SyB (Sylvain Brisse), KEH, DMA. cgMLST LIN code conceptualization and developments: MH, AC, SB. BIGSdb-Pasteur Platform maintenance: FP, BR, BB, SeB (Sebastien Bridel), SyB. Data acquisition and curation: VP, RI, CR, FP, MH, CC, SeB, ML, KW, CAY, MRP, DA. Data analyses: MH, SeB, KW, ML, CAY. PubMLST/BIGSdb platform software development: KAJ, MCJM. Pathogenwatch platform maintenance and software development: DMA, CAY, SD. Data visualization: FP, MH, KW, ML, SeB, CC. Writing of first draft: SyB, FP, MH. All authors contributed to, and approved, the final version of the manuscript.

Ethical statements

Not relevant.

Conflicts of interests

The authors declare no conflict of interest.

References

Aanensen, D.M., Spratt, B.G., 2005. The multilocus sequence typing network: mlst.net. Nucleic Acids Res 33, W728-33.

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol 62, 53–70.

Achtman, M., Zhou, Z., Charlesworth, J., Baxter, L., 2022. EnteroBase: hierarchical clustering of 100 000s of bacterial genomes into species/subspecies and populations. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 377, 20210240. https://doi.org/10.1098/rstb.2021.0240

Argimón, S., David, S., Underwood, A., Abrudan, M., Wheeler, N.E., Kekre, M., Abudahab, K., Yeats, C.A., Goater, R., Taylor, B., Harste, H., Muddyman, D., Feil, E.J., Brisse, S., Holt, K., Donado-Godoy, P., Ravikumar, K.L., Okeke, I.N., Carlos, C., Aanensen, D.M., NIHR Global Health Research Unit on Genomic Surveillance of Antimicrobial Resistance, Fabian Bernal, J., Arevalo, A., Fernanda Valencia, M., Osma Castro, E.C.D., Nagaraj, G., Shamanna, V., Govindan, V., Prabhu, A., Sravani, D., Shincy, M.R., Rose, S., Ravishankar, K.N., Oaikhena, A.O., Afolayan, A.O., Ajiboye, J.J., Ewomazino Odih, E., Lagrada, M.L., Macaranas, P.K.V., Olorosa, A.M., Gayeta, J.M., Masim, M.A.L., Herrera, E.M., Molloy, A., Stelling, J., 2021. Rapid Genomic Characterization and Global Surveillance of *Klebsiella* Using Pathogenwatch. Clin. Infect. Dis. 73, S325–S335. https://doi.org/10.1093/cid/ciab784

Bialek-Davenet, S., Criscuolo, A., Ailloud, F., Passet, V., Jones, L., Delannoy-Vieillard, A.-S., Garin, B., Le Hello, S., Arlet, G., Nicolas-Chanoine, M.-H., Decré, D., Brisse, S., 2014. Genomic definition of hypervirulent and multidrug-resistant Klebsiella pneumoniae clonal groups. Emerg. Infect. Dis. 20, 1812–1820. https://doi.org/10.3201/eid2011.140206

Bougnoux, M.-E., Aanensen, D.M., Morand, S., Théraud, M., Spratt, B.G., d’Enfert, C., 2004. Multilocus sequence typing of Candida albicans: strategies, data exchange and applications. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 4, 243–252. https://doi.org/10.1016/j.meegid.2004.06.002

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., Harris, S.R., 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 43, e15. https://doi.org/10.1093/nar/gku1196

David, S., Reuter, S., Harris, S.R., Glasner, C., Feltwell, T., Argimon, S., Abudahab, K., Goater, R., Giani, T., Errico, G., Aspbury, M., Sjunnebo, S., EuSCAPE Working Group, ESGEM Study Group, Feil, E.J., Rossolini, G.M., Aanensen, D.M., Grundmann, H., 2019. Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. Nat. Microbiol. 4, 1919–1929. https://doi.org/10.1038/s41564-019-0492-8

Deleo, F.R., Chen, L., Porcella, S.F., Martens, C.A., Kobayashi, S.D., Porter, A.R., Chavda, K.D., Jacobs, M.R., Mathema, B., Olsen, R.J., Bonomo, R.A., Musser, J.M., Kreiswirth, B.N., 2014. Molecular dissection of the evolution of carbapenem-resistant multilocus sequence type 258 Klebsiella pneumoniae. Proc. Natl. Acad. Sci. U. S. A. 111, 4988–4993. https://doi.org/10.1073/pnas.1321364111

Feil, E.J., 2004. Small change: keeping pace with microevolution. Nat Rev Microbiol 2, 483–95.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57, 81–91.

Hennart, M., Guglielmini, J., Bridel, S., Maiden, M.C.J., Jolley, K.A., Criscuolo, A., Brisse, S., 2022. A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of Klebsiella pneumoniae Strains. Mol. Biol. Evol. 39, msac135. https://doi.org/10.1093/molbev/msac135

Hetland, M.A.K., Hawkey, J., Bernhoff, E., Bakksjø, R.-J., Kaspersen, H., Rettedal, S.I., Sundsfjord, A., Holt, K.E., Löhr, I.H., 2023. Within-patient and global evolutionary dynamics of Klebsiella pneumoniae ST17. Microb. Genomics 9, mgen001005. https://doi.org/10.1099/mgen.0.001005

International Code of Nomenclature of Prokaryotes, 2019. . Int. J. Syst. Evol. Microbiol. 69, S1–S111. https://doi.org/10.1099/ijsem.0.000778

Jolley, K.A., Bray, J.E., Maiden, M.C.J., 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 3, 124. https://doi.org/10.12688/wellcomeopenres.14826.1

Jolley, K.A., Bray, J.E., Maiden, M.C.J., 2017. A RESTful application programming interface for the PubMLST molecular typing and genome databases. Database 2017, bax060. https://doi.org/10.1093/database/bax060

Konings, F., Perkins, M.D., Kuhn, J.H., Pallen, M.J., Alm, E.J., Archer, B.N., Barakat, A., Bedford, T., Bhiman, J.N., Caly, L., Carter, L.L., Cullinane, A., de Oliveira, T., Druce, J., El Masry, I., Evans, R., Gao, G.F., Gorbalenya, A.E., Hamblion, E., Herring, B.L., Hodcroft, E., Holmes, E.C., Kakkar, M., Khare, S., Koopmans, M.P.G., Korber, B., Leite, J., MacCannell, D., Marklewitz, M., Maurer-Stroh, S., Rico, J.A.M., Munster, V.J., Neher, R., Munnink, B.O., Pavlin, B.I., Peiris, M., Poon, L., Pybus, O., Rambaut, A., Resende, P., Subissi, L., Thiel, V., Tong, S., van der Werf, S., von Gottberg, A., Ziebuhr, J., Van Kerkhove, M.D., 2021. SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse. Nat. Microbiol. 6, 821–823. https://doi.org/10.1038/s41564-021-00932-w

Konstantinidis, K.T., Tiedje, J.M., 2005. Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187, 6258–64.

Lam, M.M.C., Holt, K.E., Wyres, K.L., 2023. Comment on: MDR carbapenemase-producing Klebsiella pneumoniae of the hypervirulence-associated ST23 clone in Poland, 2009-19. J. Antimicrob. Chemother. 78, 1132–1134. https://doi.org/10.1093/jac/dkad028

Lam, M.M.C., Wick, R.R., Judd, L.M., Holt, K.E., Wyres, K.L., 2022. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species complex. Microb. Genomics 8. https://doi.org/10.1099/mgen.0.000800

Lam, M.M.C., Wick, R.R., Watts, S.C., Cerdeira, L.T., Wyres, K.L., Holt, K.E., 2021. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat. Commun. 12, 4188. https://doi.org/10.1038/s41467-021-24448-3

Lam, M.M.C., Wyres, K.L., Duchêne, S., Wick, R.R., Judd, L.M., Gan, Y.-H., Hoh, C.-H., Archuleta, S., Molton, J.S., Kalimuddin, S., Koh, T.H., Passet, V., Brisse, S., Holt, K.E., 2018. Population genomics of hypervirulent Klebsiella pneumoniae clonal-group 23 reveals early emergence and rapid global dissemination. Nat. Commun. 9, 2703. https://doi.org/10.1038/s41467-018-05114-7

Maiden, M.C., 2006. Multilocus sequence typing of bacteria. Annu Rev Microbiol 60, 561–88.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U A 95, 3140–5.

Maiden, M.C., van Rensburg, M.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., McCarthy, N.D., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 11, 728–36. https://doi.org/10.1038/nrmicro3093

Marakeby, H., Badr, E., Torkey, H., Song, Y., Leman, S., Monteil, C.L., Heath, L.S., Vinatzer, B.A., 2014. A system to automatically classify and name any individual genome-sequenced organism independently of current biological classification and nomenclature. PloS One 9, e89142. https://doi.org/10.1371/journal.pone.0089142

Martin, M.J., Corey, B.W., Sannio, F., Hall, L.R., MacDonald, U., Jones, B.T., Mills, E.G., Harless, C., Stam, J., Maybank, R., Kwak, Y., Schaufler, K., Becker, K., Hübner, N.-O., Cresti, S., Tordini, G., Valassina, M., Cusi, M.G., Bennett, J.W., Russo, T.A., McGann, P.T., Lebreton, F., Docquier, J.-D., 2021. Anatomy of an extensively drug-resistant Klebsiella pneumoniae outbreak in Tuscany, Italy. Proc. Natl. Acad. Sci. U. S. A. 118, e2110227118. https://doi.org/10.1073/pnas.2110227118

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., Gilpin, B., Smith, A.M., Man Kam, K., Perez, E., Trees, E., Kubota, K., Takkinen, J., Nielsen, E.M., Carleton, H., FWD-NEXT Expert Panel, 2017. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull. 22. https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544

Paradis, E., Claude, J., Strimmer, K., 2004. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412

Prim, R.C., 1957. Shortest connection networks and some generalizations. Bell Syst Tech J 1389–1401.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Revell, L.J., 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ 12, e16505. https://doi.org/10.7717/peerj.16505

Rodrigues, C., Desai, S., Passet, V., Gajjar, D., Brisse, S., 2022. Genomic evolution of the globally disseminated multidrug-resistant Klebsiella pneumoniae clonal group 147. Microb. Genomics 8. https://doi.org/10.1099/mgen.0.000737

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Struelens, M.J., Brisse, S., 2013. From molecular to genomic epidemiology: transforming surveillance and control of infectious diseases. Euro Surveill 18, 20386.

Struelens, M.J., De Gheldre, Y., Deplano, A., 1998. Comparative and library epidemiological typing systems: outbreak investigations versus surveillance systems. Infect Control Hosp Epidemiol 19, 565–9.

van Belkum, A., Tassios, P.T., Dijkshoorn, L., Haeggman, S., Cookson, B., Fry, N.K., Fussing, V., Green, J., Feil, E., Gerner-Smidt, P., Brisse, S., Struelens, M., 2007. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect 13 Suppl 3, 1–46.

Vinatzer, B.A., Tian, L., Heath, L.S., 2017. A proposal for a portal to make earth's microbial diversity easily accessible and searchable. Antonie Van Leeuwenhoek 110, 1271–1279. https://doi.org/10.1007/s10482-017-0849-z

Wang, J., Feng, Y., Zong, Z., 2023. The Origins of ST11 KL64 Klebsiella pneumoniae: a Genome-Based Study. Microbiol. Spectr. 11, e0416522. https://doi.org/10.1128/spectrum.04165-22

Wyres, K.L., Hawkey, J., Hetland, M.A.K., Fostervold, A., Wick, R.R., Judd, L.M., Hamidian, M., Howden, B.P., Löhr, I.H., Holt, K.E., 2019. Emergence and rapid global dissemination of CTX-M-15-associated Klebsiella pneumoniae strain ST307. J. Antimicrob. Chemother. 74, 577–581. https://doi.org/10.1093/jac/dky492

Wyres, K.L., Lam, M.M.C., Holt, K.E., 2020. Population genomics of Klebsiella pneumoniae. Nat. Rev. Microbiol. 18, 344–359. https://doi.org/10.1038/s41579-019-0315-1

Wyres, K.L., Wick, R.R., Gorrie, C., Jenney, A., Follador, R., Thomson, N.R., Holt, K.E., 2016. Identification of Klebsiella capsule synthesis loci from whole genome data. Microb. Genomics 2, e000102. https://doi.org/10.1099/mgen.0.000102

Yeo, M., Mauricio, I.L., Messenger, L.A., Lewis, M.D., Llewellyn, M.S., Acosta, N., Bhattacharyya, T., Diosque, P., Carrasco, H.J., Miles, M.A., 2011. Multilocus sequence typing (MLST) for lineage assignment and high resolution diversity studies in Trypanosoma cruzi. PLoS Negl. Trop. Dis. 5, e1049. https://doi.org/10.1371/journal.pntd.0001049

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y., 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36. https://doi.org/10.1111/2041-210X.12628

Zhou, K., Xiao, T., David, S., Wang, Q., Zhou, Y., Guo, L., Aanensen, D., Holt, K.E., Thomson, N.R., Grundmann, H., Shen, P., Xiao, Y., 2020. Novel Subclone of Carbapenem-Resistant Klebsiella pneumoniae Sequence Type 11 with Enhanced Virulence and Transmissibility, China. Emerg. Infect. Dis. 26, 289–297. https://doi.org/10.3201/eid2602.190594

Zhou, K., Xue, C.-X., Xu, T., Shen, P., Wei, S., Wyres, K.L., Lam, M.M.C., Liu, J., Lin, H., Chen, Y., Holt, K.E., BRICS Working Group, Xiao, Y., 2023. A point mutation in recC associated with subclonal replacement of carbapenem-resistant Klebsiella pneumoniae ST11 in China. Nat. Commun. 14, 2464. https://doi.org/10.1038/s41467-023-38061-z

Zhou, Z., Charlesworth, J., Achtman, M., 2021. HierCC: A multi-level clustering scheme for population assignments based on core genome MLST. Bioinforma. Oxf. Engl. btab234. https://doi.org/10.1093/bioinformatics/btab234