



HAL
open science

A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of *Klebsiella pneumoniae* Strains

Melanie Hennart, Julien Guglielmini, Sébastien Bridel, Martin C J Maiden, Keith A. Jolley, Alexis Criscuolo, Sylvain Brisse

► **To cite this version:**

Melanie Hennart, Julien Guglielmini, Sébastien Bridel, Martin C J Maiden, Keith A. Jolley, et al.. A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of *Klebsiella pneumoniae* Strains. *Molecular Biology and Evolution*, 2022, 39 (7), pp.msac135. 10.1093/molbev/msac135 . pasteur-03759200

HAL Id: pasteur-03759200

<https://pasteur.hal.science/pasteur-03759200v1>

Submitted on 23 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 **A dual barcoding approach to bacterial strain nomenclature:**
2 **Genomic taxonomy of *Klebsiella pneumoniae* strains**

3

4 Melanie Hennart ^{a,b}, Julien Guglielmini ^c, Sébastien Bridel ^a, Martin C.J. Maiden ^d, Keith A. Jolley ^d, Alexis
5 Criscuolo ^c and Sylvain Brisse ^a

6

7 **Affiliations:**

8 ^a Institut Pasteur, Université Paris Cité, Biodiversity and Epidemiology of Bacterial Pathogens, F-75015,
9 Paris, France

10 ^b Sorbonne Université, Collège doctoral, F-75005 Paris, France

11 ^c Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France

12 ^d Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford, OX1 3SZ, United Kingdom

13 ***Correspondence:**

14 Sylvain Brisse. Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, F-75724 Paris,
15 France. E-mail: sylvain.brisse@pasteur.fr; Tel: +33 1 45 68 83 34

16 **Keywords:** genomic classification, strain nomenclature, microevolution, pathogen tracking, genomic
17 library, international harmonization

18 **Running Title:** *Klebsiella* strain taxonomy

19 **Abstract**

20 Sublineages within microbial species can differ widely in their ecology and pathogenicity, and their
21 precise definition is important in basic research and for industrial or public health applications. Widely
22 accepted strategies to define sublineages are currently missing, which confuses communication in
23 population biology and epidemiological surveillance.

24 Here we propose a broadly applicable genomic classification and nomenclature approach for bacterial
25 strains, using the prominent public health threat *Klebsiella pneumoniae* as a model. Based on a 629-
26 gene core genome multilocus sequence typing (cgMLST) scheme, we devised a dual barcoding system
27 that combines multilevel single linkage (MLSL) clustering and life identification numbers (LIN).
28 Phylogenetic and clustering analyses of >7,000 genome sequences captured population structure
29 discontinuities, which were used to guide the definition of 10 infra-specific genetic dissimilarity
30 thresholds. The widely used 7-gene multilocus sequence typing (MLST) nomenclature was mapped
31 onto MLSL sublineages (threshold: 190 allelic mismatches) and clonal group (threshold: 43) identifiers
32 for backwards nomenclature compatibility. The taxonomy is publicly accessible through a community-
33 curated platform (<https://bigsd.b.pasteur.fr/klebsiella>), which also enables external users' genomic
34 sequences identification.

35 The proposed strain taxonomy combines two phylogenetically informative barcode systems that
36 provide full stability (LIN codes) and nomenclatural continuity with previous nomenclature (MLSL). This
37 species-specific dual barcoding strategy for the genomic taxonomy of microbial strains is broadly
38 applicable and should contribute to unify global and cross-sector collaborative knowledge on the
39 emergence and microevolution of bacterial pathogens.

40 Introduction

41 Taxonomy is a foundation of biology that entails the classification, nomenclature, and identification of
42 biological objects (Cowan 1965). Although the Linnaean system is organized into taxonomic ranks
43 down to the level of species (Sneath 1992), sublineages within microbial species can diversify as
44 independently evolving lineages that persist over long periods of time, (Selander and Levin 1980) and
45 the broad microbial species definition and horizontal gene transfer of accessory genes underlie
46 extensive strain heterogeneity of phenotypes with ecological, medical or industrial relevance (Hacker
47 and Kaper 2000; Lan and Reeves 2001; Feil 2004; Konstantinidis and Tiedje 2005). Nevertheless, strain-
48 level diversity is overlooked by current prokaryotic taxonomy.

49 Most attempts to develop and maintain microbial strain taxonomies aimed at facilitating
50 epidemiological surveillance and outbreak detection (Maiden et al. 1998; van Belkum et al. 2007;
51 Maiden et al. 2013). Although local epidemiology can rely on vernacular type designations, the benefits
52 of unified nomenclatures of sublineages for large-scale epidemiology and population biology were
53 recognized early (Struelens, De Gheldre, and Deplano 1998). By far the most successful taxonomic
54 system of microbial strains is the multilocus sequence typing (MLST) approach (Maiden et al. 1998;
55 Achtman et al. 2012). This highly reproducible and portable nomenclature system has been extensively
56 used for studies of population biology and public health surveillance of bacterial pathogens (Keith A.
57 Jolley, Bray, and Maiden 2018). Core genome MLST (cgMLST) extends the advantages of the MLST
58 approach at the genomic scale (K. A. Jolley and Maiden 2010; Maiden et al. 2013) and provides strain
59 discrimination at much finer scales.

60 Strain classification, based either on cgMLST or on nucleotide polymorphisms, can be achieved by using
61 several clustering thresholds simultaneously, leading to a succession of group identifiers ('barcodes')
62 that provide relatedness information at increasing levels of phylogenetic depth (Maiden et al. 2013;
63 Moura et al. 2016). This approach was recently formalized as hierarchical clustering (HierCC, based on
64 cgMLST) (Zhou, Charlesworth, and Achtman 2021) and as the 'single nucleotide polymorphism (SNP)
65 address' (Dallman et al. 2018) , based on single linkage classifications; here we generically refer to
66 these approaches as MultiLevel Single Linkage (MLSL). Unfortunately, the single linkage clustering may
67 result in the fusion of preexisting groups as additional genomes are introduced, due to the possibility
68 of new genomes being less distant than the threshold, from two distinct groups. This approach thus
69 suffers from instability, which led HierCC inventors to instead use *ad-hoc* group attribution rules after
70 an initial single linkage classification (Zhou, Charlesworth, and Achtman 2021).

71 An alternative approach, the Life Identification Number (LIN) encoding, was proposed by Vinatzer and
72 colleagues (Vinatzer, Tian, and Heath 2017; Tian et al. 2020): a multi-position numerical code is
73 assigned to each genome based on its similarity with the closest genome already encoded. An
74 attractive property of this procedure is that LIN codes are definitive, *i.e.*, not affected by subsequent
75 additions of genomes, as they are attributed to individual genomic sequences rather than to groups.
76 However, in the current implementation of LIN codes the similarity between genomes is estimated
77 using Average Nucleotide Identity (ANI), which may be imprecise for nearly identical strains.

78 Here, we present a strain classification, naming and identification system for bacterial strains, which is
79 based on cgMLST and combines the MLSL and LIN code approaches. We took as a model the
80 *Klebsiella pneumoniae* species complex, a genetically and ecologically highly diverse bacterial group
81 that causes a wide range of infections in humans and animals (Brisse, Grimont, and Grimont 2006;
82 Wyres, Lam, and Holt 2020). Given its extensive diversity and fast evolutionary dynamics,
83 *K. pneumoniae* is a challenging model for the development of a genomic taxonomy of strains.
84 Moreover, the rapid emergence and global dissemination of multidrug resistance in *K. pneumoniae*,
85 sometimes combined with high virulence, (Bialek-Davenet et al. 2014; Wyres et al. 2020) have created
86 a pressing need for an efficient *K. pneumoniae* strain definition and tracking system.

87 Results

88 *Genome-based phylogenetic structure of the K. pneumoniae species complex (KpSC)*

89 The deep phylogenetic structure of the *K. pneumoniae* (Kp) species complex (**Figure 1**) reflects the
90 previously recognized seven major phylogroups, Kp1 to Kp7 (Brisse and Verhoef 2001; Fevre et al.
91 2005; Blin et al. 2017; Long et al. 2017; Rodrigues et al. 2019; Wyres, Lam, and Holt 2020). The most
92 represented phylogroup (91.7%; n=6,476) is Kp1, *i.e.*, *K. pneumoniae sensu stricto* (**Table 1**), and its
93 phylogenetic structure (**Figure 2**) revealed a multitude of sublineages (note that below, we define
94 sublineages and clonal groups in a stricter sense in paragraph “Definition of classification thresholds
95 for phylogroups, sublineages and clonal groups”). There were multiple closely-related isolates within
96 some sublineages, most prominently within a sublineage comprising genomes with 7-gene MLST
97 identifiers ST258, ST11, ST512 (**Figure 2**), which represented more than a third (33.4%) of the Kp1
98 dataset. The abundance of this sublineage (and a few others, such as ST23) reflected the clinical
99 microbiology focus on multidrug resistant or hypervirulent isolates (Bowers et al. 2015; Struve et al.
100 2015; Lam et al. 2018; Wyres et al. 2020). The phylogenetic structure within other *K. pneumoniae*
101 phylogroups also revealed a multitude of distinct sublineages but no predominant ones, and medically
102 important lineages in these phylogroups are yet to be recognized.

103 *K. pneumoniae* strains can recombine large sections of their chromosome (Chen et al. 2014; Wyres et
104 al. 2015). Large recombination events were detected in 1.9% (138/7,198) genomes (based on their
105 cgMLST profiles) and involved the phylogroups Kp1, Kp2 and Kp4 (supplementary appendix: Detection
106 of hybrids; **Figure S1**; **Table S1**; **Table S2**). The phylogenetic impact of large-scale recombination is
107 illustrated on **Figure 1**, with ‘hybrids’ occurring on misleadingly long branches.

108

109 *cgMLST analysis of the K. pneumoniae species complex*

110 A previously defined core genome MLST (named scgMLST, with 634 loci) scheme (Bialek-Davenet et
111 al. 2014) was updated (**Table S3**) and defined as scgMLSTv2 (with 629 loci, as five of the original ones
112 were removed; see Methods). cgMLST allelic profiles were then determined for 7,433 genomic
113 sequences (including 45 reference sequences; **Figure S2**). The mean number of missing alleles per
114 profile was 8 (1.2%; standard deviation: 25; 4.0%), and most (7,198; 96.8%) isolates had a cgMLST
115 profile with fewer than 30 (4.8%) missing alleles. Missing allele proportions did not vary significantly
116 among phylogroups (**Table 1**). The transcription-repair coupling factor *mfd* gene was atypical, with 778
117 alleles and an average allele size of 3,447 nucleotides (nt); for the other loci, the number of distinct

118 alleles varied from 8 to 626 (median: 243), and was strongly associated with locus size (range: 123 to
119 2,826 nt; median: 758 nt; **Figure S3**). Locus-by-locus recombination analyses detected evidence of
120 intra-gene recombination (PHI test; 5% *p*-value significance) in half of the loci (318/629; 50.6%) and
121 these exhibited more alleles than non-recombining ones (**Table S3; Figure S3**).

122 The distribution of pairwise allelic mismatch proportions among non-hybrid cgMLST allelic profiles was
123 discontinuous (**Figure 3**), with four major modes centered around values 99.7% (627 mismatches; *i.e.*,
124 0.3% similarity), 82.7% (520 mismatches; 17.3%), 12.4% (79 mismatches; 87.6%) and 2.0% (13
125 mismatches; 98%). Average nucleotide identity (ANI) values (**Figure S4**) varied from 92.8% to 100%,
126 with two first modes at 93.5% and 95.5%, composed of inter-phylogroup strain comparisons. The
127 corresponding genome pairs typically had only \approx 2% cgMLST similarity. In turn, whereas the range of
128 ANI values was only 98% to 100% for intra-species pairs, their cgMLST similarities occupied the much
129 broader 5%-100% range.

130 The 627-mismatch mode corresponded mostly to pairs of strains belonging to distinct species of the
131 KpSC (**Figure 3; Figure S4**), while a minor peak centered on 591 mismatches (**Figure S5**) corresponded
132 to comparisons between subspecies of *K. quasipneumoniae* and *K. variicola* (Kp2 and Kp4, and Kp3 and
133 Kp5, respectively; **Figure S4**). Whereas the 520-mismatch mode corresponded to inter-ST comparisons
134 in 99.9% cases, the 13-mismatch mode was largely dominated by comparisons of cgMLST profiles with
135 the same ST (68.2%; pairs of genomes within 402 distinct STs) or of single-locus variants (SLV; 30.8%).
136 Finally, the 79-mismatch mode comprised a large proportion (48.0%) of ST258-ST11 comparisons and
137 other comparisons of atypically closely-related STs (**Figure S5**).

138

139 ***Definition of classification thresholds for phylogroups, sublineages and clonal groups***

140 To determine optimal allelic mismatch thresholds that would reflect the KpSC population structure,
141 the consistency and stability properties of single linkage clustering groups were assessed for every
142 threshold value *t* from 1 to 629 allelic mismatches. The consistency (silhouette) coefficient S_t had a
143 plateau of optimal values in the range corresponding to 118/629 (18.8%) to 355/629 (56.4%) allelic
144 mismatches (**Figure 3, blue curve**). Analysis of the robustness to subsampling (W_t ; based on an
145 adjusted Wallace coefficient; **Figure 3, green curve**) identified several ranges of allelic mismatch
146 threshold values that were associated to maximal stability.

147 The above analyses led us to propose four deep classification levels. The two first thresholds, 610 and
148 585 allelic mismatches, enable species and subspecies separations, respectively. We next defined a

149 threshold of 190 allelic mismatches, corresponding to the optimal combination of consistency and
150 stability coefficients S_t and W_t . The single linkage clustering based on this threshold created 705 groups,
151 which we here define as ‘sublineages’ (SL). By design, this threshold separated into distinct groups, the
152 pairs of cgMLST profiles corresponding to the major mode (at 520 mismatches), *i.e.*, the majority of
153 genomes that have distinct STs within phylogroups. Finally, a threshold of 43 allelic mismatches was
154 defined to separate genome pairs of the 79-mismatch mode. This value corresponded to local optima
155 of both S_t and W_t coefficients. Interestingly, this last threshold value was also located in the optimal
156 range of compatibility with the classical 7-gene ST definitions (Rand index $R_t \geq 0.70$ was observed for
157 $10 \leq t \leq 51$). The use of this threshold resulted in 1,147 groups, which we propose to define as ‘clonal
158 groups’ (CG).

159 Overall, approximately half (547/1,147; 47.7%) of the CGs corresponded one-to-one with the
160 sublineage level (**Table S4**): 77.6% (547/705) sublineages contained a single CG, whereas 158 (22.4%)
161 sublineages comprised at least two clonal groups (**Table S5; Figure 4; Figure S6**). Overall, CG
162 compatibility with classical ST classification was high (*i.e.*, $R_t = 0.72$, whereas it was only 0.50 for
163 sublineages).

164 The distribution of pairwise allelic mismatch values that involved hybrid genomes showed an
165 additional peak around 39 shared alleles (*i.e.*, 590 allelic mismatches; **Figure S7**). Therefore, these
166 inter-phylogroup hybrids were placed into distinct partitions at the 585-mismatch level. However, as
167 some of these hybrid genomes diverged by fewer than 585 allelic mismatches from two distinct
168 phylogroups at the same time, they would cause fusion of phylogroup partitions upon single linkage
169 clustering. To highlight the impact of this phenomenon, hybrids were first filtered out, and next
170 incorporated in a second single linkage clustering step (supplementary appendix; **Figure S8**).

171

172 ***Phylogenetic compatibility of sublineages and clonal groups***

173 To estimate the congruence of classification groups with phylogenetic relationships among genomic
174 sequences, we quantified the proportion of monophyletic (single ancestor, exclusive group),
175 paraphyletic (single ancestor, non-exclusive group) and polyphyletic (two or more distinct ancestors)
176 groups. Regarding 7-gene MLST, 6,985 (98.9%) genomes had a defined ST, *i.e.*, an allele was called for
177 each of the seven genes. Of the 992 distinct STs, 396 were non-singleton STs (*i.e.*, comprised at least
178 two isolates). Of these, 286 (72.2%) were monophyletic, nine were paraphyletic (2.3%) and 101 (25.5%)
179 were polyphyletic. The monophyletic STs comprised only 22% of all genomes in non-singleton STs.

180 Regarding cgMLST-based classification, there were five and seven partitions at 610 and 585 allelic
181 mismatch levels, respectively, and 100% of these were monophyletic. Among the 705 distinct
182 sublineages (SLs), 317 (45.0%) were non-singleton, and most (310; 97.8%) of these were monophyletic
183 (**see Figure 2 for Kp1**); only three (0.9%) were paraphyletic, and four (1.3%) were polyphyletic
184 (**Table S4**). The monophyletic SLs comprised a large majority (5,961/6,672; 89.3%) of genomes in non-
185 singleton STs.

186 Finally, 396 out of 1,147 (34.5%) clonal groups (CGs) were non-singleton; most (362; 91.4%) were
187 monophyletic (**Figure 2**), whereas eight (2.0%) were paraphyletic, and 26 (6.6%) were polyphyletic
188 (**Table S5**). Monophyletic CGs comprised nearly half (3,030; 48.0%) of the genomes in the non-
189 singleton CGs, whereas 3,224 (51.1%) were in polyphyletic groups, mostly in CG258, CG340 and CG15.

190

191 ***Definition of shallow-level classification thresholds for Klebsiella epidemiology***

192 Although the scgMLSTv2 scheme comprises only 629 loci, or ~10% of a typical *K. pneumoniae* genome
193 length (512,856 nt out of 5,248,520 in the NTUH-K2044 genome), shallow-level classifications of
194 genomic sequences might be useful for tentative outbreak delineation and epidemiological
195 surveillance purposes, by ruling-out outliers. To provide flexible case cluster definitions, we classified
196 KpSC cgMLST profiles using thresholds of 0, 1, 2, 4, 7 and 10 scgMLSTv2 allelic mismatches. Together
197 with the four higher levels, the MSLSL nomenclature therefore comprises 10 classification levels in total.
198 The classification groups corresponding to the 0-mismatch threshold correspond to groups of cgST
199 profiles that only differ by missing data. We observed that profiles of isolates involved in previously
200 reported KpSC outbreaks generally differed by no or 1 mismatch, with a maximum of five allelic
201 mismatches (**Table S6**; **Table S7**), indicating that this classification approach may be useful for genomic
202 surveillance and outbreak identification purposes.

203

204 ***Inheritance of the 7-gene ST identifiers into the cgMLST classification, and characteristics of main*** 205 ***sublineages (SLs) and clonal groups (CGs)***

206 To attribute SL and CG identifiers that corresponded maximally to the widely adopted 7-gene ST
207 identifiers, we developed an inheritance algorithm to map MLST identifiers onto SL and CG partitions
208 (see supplementary appendix: Nomenclature inheritance algorithm). Of the 705 SLs, most (683; 96.9%)
209 were named by inheritance and this was the case for 879 (76.6%) of the 1,047 CGs (**Table S4**). The
210 resulting correspondence of cgMLST partitions with classical MLST was evident for the major groups

211 **(Figure 4; Figure S6)**. For instance, the multidrug resistant SL258 comprised isolates belonging to MLST
212 sequence types ST258, ST11, ST512, ST340, ST437 and 25 other STs. SL258 consisted of 16 distinct CGs,
213 of which the four most frequent were defined as CG258 (61.2%), CG340 (17.8%), CG11 (17.3%) and
214 CG3666 (2.8%) **(Figure 4)**. When compared to 7-gene MLST, most isolates of CG258 were ST258 (75.6%)
215 or ST512 (22.6%), whereas CG11 mostly comprised ST11 genomes (98.0%). In turn, CG340 included a
216 large majority of ST11 genomes (61.8%) and only 20.0% ST340 genomes, and was named CG340 rather
217 than CG11 because CG11 was already attributed. Likewise, the majority (83/86; 96.5%) of ST23
218 genomes, which are associated with pyogenic liver abscess (Lam et al. 2018), were classified into SL23,
219 which itself consisted mainly (84/90, 93.3%) of ST23 genomes **(Figure 4)**. The well-recognized emerging
220 multidrug resistant KpSC populations of ST15, ST101, ST147 and ST307 each corresponded largely to a
221 single SL and CG **(Figure S6)**.

222 The frequency of detection of virulence and antimicrobial resistance genes differed among the main
223 SLs and CGs **(Figure 5; Figure S9)**. As expected (Lam et al. 2021), SL23 (median virulence score of 5) and
224 SL86 (median score 3) were prominent ‘hypervirulent’ sublineages, and they were largely lacking
225 resistance genes. In contrast, a majority of strains from SLs 258, 147, 101, 307 and 37, as well as a large
226 number of other SLs, had a resistance score of 2 or more, indicative of BLSE/carbapenemases, but
227 these had modest virulence scores **(Figure S9)**. SL231 genomes stood out as combining high virulence
228 and resistance scores. In some cases, CGs within single major SLs had contrasted virulence and
229 resistance gene contents **(Figure 6)**.

230

231 ***Development and implementation of a cgMLST-based LIN code system***

232 Following the principle of the LIN code system, initially proposed based on the ANI similarity (Marakeby
233 et al. 2014), we defined a cgMLST-based LIN (cgLIN) code approach. As LIN coding is performed
234 sequentially, we first explored the impact on the resulting partitioning of cgMLST profiles, of the order
235 in which genomes are assigned. We confirmed that the number of partitions (hence their content too)
236 varied according to input order **(Figure S10)**. However, we established that the order of genomes
237 determined by the traversal of a Minimum Spanning tree (MStree) (Prim 1957, 57) naturally induces a
238 LIN encoding order that is optimal, *i.e.*, most parsimonious with respect to the number of identifiers
239 generated at each position of the code (see supplementary appendix). Using this MStree traversal
240 strategy, we defined cgLIN codes for the 7,060 non-hybrid genomes (as a first step), resulting in 4,889
241 distinct cgLIN codes.

242 Furthermore, cgLIN codes can be displayed in the form of a prefix tree (**Figure 5**), which largely reflects
243 the phylogenetic relationships among genomes. In addition, cgLIN code prefixes can be used to label
244 particular phylogenetic lineages (Vinatzer, Tian, and Heath 2017). For example, a single cgLIN code
245 prefix defined each phylogroup (*e.g.*, Kp1: prefix 0_0; Kp2: prefix 2_0; **Figure 5**). Likewise, a full one-
246 to-one correspondence between prefixes and SLs was observed, and almost all (99.4%) CGs also had a
247 unique prefix (**Table S4; Table S5; Figure 6**).

248

249 ***Effect of hybrid genomes incorporation on the MLSL and cgLIN codes classifications***

250 Because inter-phylogroup hybrid genomes have smaller distances to their parental phylogroups than
251 the inter-phylogroup distances resulting from vertical evolutionary events, their incorporation into the
252 MLSL classification may induce fusion of previously distinct single linkage groups. To illustrate this
253 chaining effect, the 'hybrid' genomes were included into the MLSL nomenclature in a second step, and
254 fusions of previously existing partitions we recorded; for example, at the 610 allelic mismatch
255 threshold, partitions 2 (Kp2 and Kp4) and 4 (Kp3 and Kp5) were merged with partition 1 (Kp1). At the
256 585-mismatch threshold, partitions 5 (Kp2) and 2 (Kp4) were merged with partition 1 (Kp1). At the 190-
257 mismatch threshold, only one fusion was observed, between partitions 184 (SL113) and 465 (SL1518;
258 **Figure S11**). The partitions at other thresholds were not impacted by the addition of the hybrid
259 genomes.

260 In contrast, the incorporation of hybrid genomes into the cgLIN code database left the cgLIN codes of
261 the 7,060 previous genomes entirely unaffected; there were no merging of groups, as per design of
262 the system. In particular, the seven phylogroup prefixes corresponding to species and subspecies
263 remained unaffected (**Figure S11**); instead, additional prefixes were created for the hybrid genomes
264 (**Table S1; Table S2**).

265

266 ***Implementation of the genomic taxonomy in a publicly-accessible database***

267 The MLSL nomenclature was incorporated into the Institut Pasteur *K. pneumoniae* MLST and whole
268 genome MLST databases (<https://bigsdbs.pasteur.fr/klebsiella>) under the classification scheme
269 functionality developed in BIGSdb version 1.21.0. In brief, the cgMLST profile of every isolate with
270 fewer than 30 missing scgMLSTv2 alleles were assigned to a core genome sequence type (cgST), and
271 these were next grouped into single linkage partitions for each of the 10 classification levels. For SLs
272 and CGs, a custom classification group field (named SL or CG within the system) was additionally

273 populated with identifiers inherited from 7-gene MLST. All cgMLST profiles and classification identifiers
274 are publicly available.

275 To allow users identifying *K. pneumoniae* isolates easily, a profile matching functionality was
276 developed, enabling to search for cgMLST profiles related to a query genome sequence. This was
277 implemented on the website sequence query page
278 ([https://bigsdbs.readthedocs.io/en/latest/administration.html#scheme-profile-clustering-setting-up-
279 classification-schemes](https://bigsdbs.readthedocs.io/en/latest/administration.html#scheme-profile-clustering-setting-up-classification-schemes)). This functionality returns the classification identifiers (including MLST-
280 inherited CG and SL identifiers) of the cgMLST profile that is most closely related to the query genomic
281 sequence, along with its number of mismatches compared to the closest profile.

282 cgLIN code functionality was also incorporated into BIGSdb version v1.34.0
283 ([https://bigsdbs.readthedocs.io/en/latest/administration.html#setting-up-lincode-definitions-for-
284 cgmlst-schemes](https://bigsdbs.readthedocs.io/en/latest/administration.html#setting-up-lincode-definitions-for-cgmlst-schemes)). In particular, cgST profiles can be queried by full cgLIN code or any prefix, and a
285 nomenclature can be attached to LINcode prefixes of interest (*e.g.*, SL258 is attached to prefix 0_0_1
286 and CG258 to 0_0_105_6).

287 Note that identification of users' query genomic sequences is made possible either through the BIGSdb
288 platform that underlies the cgMLST website, or externally after export of the cgMLST profiles, which
289 are publicly accessible.

290

291 Discussion

292 The existence within microbial species of sublineages with unique genotypic and phenotypic properties
293 underlines the need for infra-specific nomenclatures (Lan and Reeves 2001; Maiden et al. 1998;
294 Rambaut et al. 2020). Similar to species and higher Linnaean taxonomic ranks, a strain taxonomy
295 should: (i) recognize genetic discontinuities and capture the most relevant sublineages at different
296 phylogenetic depths; (ii) provide an unambiguous naming system for sublineages; and (iii) provide
297 identification methods for placement within the taxonomic framework. Here we developed a strain
298 taxonomy consisting of a dual naming system that is grounded in population genetics and linked to an
299 identification tool. The proposed system thus complies with the three fundamental pillars of
300 taxonomy.

301 Although 7-gene MLST has been widely adopted as a taxonomic system of KpSC strains, several
302 limitations are apparent: besides its restricted resolution, MLST identifiers do not convey phylogenetic
303 information, as a single nucleotide substitution generates a different ST with unapparent relationships
304 with its ancestor. Further, approximately half of the ST partitions were not monophyletic. The cgMLST
305 approach provides much higher resolution and phylogenetic precision (Maiden et al. 2013; Zhou,
306 Charlesworth, and Achtman 2021). Although other metrics such as whole-genome single nucleotide
307 polymorphisms (SNPs) or average nucleotide identity (ANI) can be used to classify strains (Marakeby
308 et al. 2014; Dallman et al. 2018), cgMLST presents advantages inherited from classical MLST, including
309 standardization, reproducibility, portability and the conversion of sequences into human-readable
310 allelic numbers. The high reproducibility and easy interpretation of cgMLST are two critical
311 characteristics for its adoption in epidemiological surveillance. Here, we showed that cgMLST, based
312 on 629 genes, has a much broader dynamic range than ANI when considering intra-specific variation
313 (**Figure S4**), and enables defining several hierarchical classification levels (Zhou, Charlesworth, and
314 Achtman 2021). The resolutive power of the 629-loci cgMLST scheme provides valuable genotyping
315 discrimination up to outbreak resolution and is highly consistent with whole-genome SNPs (Miro et al.
316 2020). However, to define shallower genetic structure within sublineages resulting from recent clonal
317 expansions or outbreaks, higher resolution should be sought based *e.g.*, on core gene sets of specific
318 sublineages.

319 Optimization of threshold definitions based on population structure aims at optimizing cluster stability
320 (Barker et al. 2018; Zhou, Charlesworth, and Achtman 2021). The density distribution of pairwise allelic
321 mismatch dissimilarities within *K. pneumoniae* and related species exhibited genetic discontinuities at
322 several phylogenetic depths. We took the benefit of this multimodal distribution to define optimal

323 intra-specific classification thresholds, and combined a clustering consistency coefficient (Silhouette)
324 with a newly developed strategy that evaluates cluster stability by subsampling the entire dataset. We
325 defined four classifications at phylogenetic depths that reflected natural discontinuities within the
326 population structure of *K. pneumoniae*, including the deep subdivisions of *K. pneumoniae sensu lato*.
327 The 190-mismatch sublineage level was designed to capture the numerous deep phylogenetic
328 branches within phylogroups (Bialek-Davenet et al. 2014; Holt et al. 2015). In turn, the clonal group
329 (CG) level was useful to capture the genetic structuration observed within these primary sublineages.
330 For example, the CG-level nomenclature captures the evolutionary split of CG258 and CG11 from their
331 SL258 ancestor, caused by a 1.1 Mb recombination event (Chen et al. 2014) (**Figure S5**).

332 Seven-locus MLST is a widely adopted nomenclature system, as illustrated by the widespread use of
333 ST identifiers associated with hypervirulent or multidrug resistant sublineages (*e.g.*, ‘*Klebsiella*
334 *pneumoniae* ST258’: 293 PubMed hits; ST23: 117 hits; on July 20th, 2021). Backward nomenclatural
335 compatibility is therefore critical. After applying our inheritance algorithm, most sublineages and
336 clonal groups were labeled according to the 7-gene MLST identifier of the majority of their isolates.
337 Widely adopted ST identifiers will therefore designate nearly the same strain groups within the
338 proposed genomic taxonomy of *K. pneumoniae*, which should greatly facilitate its adoption. We note
339 that the 7-gene MLST nomenclature will still have to be expanded, as this classical approach continues
340 to be widely used. However, for practical reasons, upcoming MLST and cgMLST nomenclatural
341 identifiers will be uncoupled, and we suggest that the cgMLST-based identifiers of sublineages and
342 clonal groups (rather than their ST) should be adopted as the reference nomenclature in the future.

343 Instability is a major limitation of single linkage clustering, caused by group fusion known as the
344 chaining effect (Turner and Feil 2007). This is particularly relevant over epidemiological timescales,
345 where intermediate genotypes (*e.g.*, a recent ancestor or recombinant) are often sampled (Feil 2004).
346 This issue is exacerbated in *K. pneumoniae*, where large-scale recombination may result in truly
347 intermediate genotypes (Chen et al. 2014; Holt et al. 2015), referred to as ‘hybrids’ by analogy to
348 eukaryotic biology. Here this phenomenon was illustrated through our delayed introduction into our
349 nomenclature, of 138 inter-phylogroup hybrid genomes. The merging of predefined classification
350 groups can be handled by classification versioning or *ad hoc* rules (Zhou, Charlesworth, and Achtman
351 2021), but this is indeterministic and challenging in practice.

352 To address its stability issue, we complemented the single linkage clustering approach with a fully
353 stable approach. LIN codes were proposed as a universal genome coding system (Marakeby et al. 2014;
354 Tian et al. 2020), a key feature of which is the generation of definitive genome codes that are inherently
355 stable. The original LIN code system was based on the ANI metric; here we noted that the ANI values

356 that best correspond to some of the 10 cgMLST thresholds were highly similar (**Table S8**), casting doubt
357 on the reliability of this metric for small-scale genetic distances. In addition, the ANI metric is non-
358 reciprocal and highly dependent on comparison implementations and parameters. These shortfalls
359 may yield imprecision and non-reproducibility that are particularly impactful for comparisons between
360 very similar genomes. We therefore adapted the LIN code concept to cgMLST-based similarity (*i.e.*,
361 one-complement of the allelic mismatch proportions) to classify strains into a cgLIN code system. This
362 strategy leverages the benefits of cgMLST, and introduces more intuitive shallow-level classification
363 thresholds. The multilevel similarity information embedded in MSL and cgLIN ‘barcodes’ provides a
364 human-readable snapshot of strain relationships, as nearly identical genomes have identical barcodes
365 up to a position near the right end. In contrast to single linkage clustering partitions, one important
366 limitation of LIN codes is that preexisting classification identifiers (*e.g.*, ST258) cannot be mapped onto
367 individual LIN code identifiers, because these are attributed with reference to the upper levels and are
368 set to 0 for each downstream level (Marakeby et al. 2014). However, cgLIN code prefixes may
369 represent useful labels for particular lineages.

370 The cgMLST-based nomenclatures have some limitations. First, comparison of allele numbers rather
371 than SNPs implies loss of information. In turn, this approach is advantageous to estimate evolutionary
372 relationships of closely related genomes that diverged following homologous recombination events,
373 which are common among strains within bacterial species (Feil 2004; Vos and Didelot 2009). Second,
374 MLST-based distances saturate more rapidly than SNP distances (once a locus is affected, even by a
375 single mutation, further mutations at this locus will change the allele but will not increase the allelic
376 distance), and are therefore mostly meaningful within bacterial species. Third, in contrast to the
377 original ANI-based LIN code approach (Tian et al. 2020), cgMLST-based LIN codes require prior
378 development of cgMLST schemes, which are larger, hence more powerful for strain resolution, for
379 single species. Therefore, the advantages of cgLIN codes for population genomics and epidemiological
380 questions, come at the expense of universality. Still, the dual MSL and cgLIN code approach proposed
381 here is in principle applicable to all bacterial species (or closely related groups thereof) for which large
382 representative sets of genomes are available. The cgLIN code algorithms were incorporated into
383 BIGSdb and should be readily portable to other existing cgMLST platforms such as Enterobase.
384 However, the use of genetic thresholds for an entire group of organisms may not always be meaningful,
385 depending on population genetic structure. Whereas *K. pneumoniae* shows strong structuring with
386 neat peaks and valleys of pairwise genetic distances, other species may have more fuzzy structure. In
387 the latter cases, the approach will still be applicable, but even optimally defined thresholds may be
388 less relevant biologically.

389 **Conclusions**

390 A unified nomenclature of pathogen genotypes is required to facilitate communication in the 'One
391 Health' and 'Global Health' perspectives. *K. pneumoniae* represents a rapidly growing public health
392 threat, and the availability of a common language to designate its emerging sublineages is therefore
393 highly timely. The proposed unified taxonomy of *K. pneumoniae* strains will facilitate advances on the
394 biology of its sublineages across niches, time and space, and will endow surveillance networks with the
395 capacity to efficiently monitor and control the emergence of sublineages of high public health
396 relevance.

397 Here, we propose a dual barcoding approach to bacterial strain taxonomy, which combines the
398 complementary advantages of stability provided by the cgLIN codes, with an unstable, but human
399 readable multilevel single linkage nomenclature rooted in the popular 7-gene MLST nomenclature.
400 Because they are definitive, cgLIN codes can be used for the traceability of cluster fusions that will
401 occur occasionally in the MLSL arm of the dual taxonomy (**Figure S11**). We contend the stability of
402 cgLIN codes and their use alongside MLSL approaches provide a pragmatic solution to current attempts
403 at developing genomic taxonomies of bacterial strains that are both stable and practical for human-
404 to-human communication.

405 **Material & Methods**

406 ***Definition of an updated core genome MLST (scgMLSTv2) genotyping scheme***

407 We previously defined a core genome MLST (using *strict* synteny criteria, hence name scgMLSTv1)
408 scheme of 634 highly syntenic genes (Bialek-Davenet et al. 2014). Here, we updated the scgMLSTv1
409 scheme, with the following improvements. First, two loci (KP1_2104 and *aceB*=KP1_0253) were
410 removed because they were absent or truncated in multiple strains, based on 751 high-quality
411 assemblies available in the BIGSdb-Pasteur *Klebsiella* database on October 16th, 2017 (project id 11 at
412 https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_isolates&page=projects).
413 Second, the remaining 632 loci templates were modified so that they would include the start and stop
414 codons of the corresponding coding sequence (CDS). This was not the case for all CDSs of the
415 scgMLSTv1 scheme, as some loci corresponded to internal portions of CDSs. These template
416 redefinitions were done to harmonize locus definitions across the scheme. Of note, defining loci as
417 complete CDSs also facilitates genotyping, by enabling precise identification of the extremities of novel
418 alleles, through the search of the corresponding start and stop codons. As a result of these locus
419 template extensions, three additional loci (*yraR*, *rnt* and KP1_1655) had to be removed because they
420 were called in a low proportion of the above 751 genomes. The resulting 629 scgMLSTv2 genes have a
421 summed length of 512,856 nt (9.8% of the genome of reference strain NTUH-K2044), as compared to
422 507,512 nt (9.7%) for the corresponding loci in scgMLSTv1.

423

424 ***Definition of a genomic sequence dataset of 7,060 isolates with cgMLST profiles***

425 The *K. pneumoniae* species complex (KpSC) comprises seven phylogroups that have been given
426 taxonomic status in the prokaryotic nomenclature: *K. pneumoniae* subsp. *pneumoniae* (Kp1, also
427 known as *K. pneumoniae sensu stricto*), *K. quasipneumoniae* subsp. *quasipneumoniae* (Kp2), *K.*
428 *variicola* subsp. *variicola* (Kp3), *K. quasipneumoniae* subsp. *similipneumoniae* (Kp4), *K. variicola* subsp.
429 *tropica* (Kp5), '*K. quasivariicola*' (Kp6) and *K. africana* (Kp7) (Rodrigues et al. 2019). We retrieved all
430 KpSC genomes from the GenBank assembly repository on March 15th, 2019, corresponding to 8,125
431 assemblies. We then chose high-quality assemblies by excluding draft genomes: (i) containing more
432 than 1,000 contigs of size >200 nt; (ii) for which the average nucleotide identity (ANI) values (estimated
433 using FastANI v1.1) were < 96% against every reference strain of the taxonomic diversity of the SC
434 (Rodrigues et al., 2019; **Table S9**); (iii) of size ≤ 4.5Mb or ≥ 6.5 Mb; and (iv) with G+C% content >59%.
435 The data of each criterion per strain are shown in **Table S1**. The three last criteria excluded possible
436 contamination or non-KpSC genomes (**Figure S2**).

437 The resultant 7,388 ‘high-quality’ draft genomes (**Table S2**) were scanned for scgMLSTv2 alleles, using
438 the BLASTN algorithm, as implemented in the BIGSdb platform (Keith A. Jolley, Bray, and Maiden 2018;
439 K. A. Jolley and Maiden 2010), with 90% identity, 90% length coverage, word size 30, with type alleles
440 only (as defined below). After this step, 235 profiles were excluded because they had more than 30
441 missing alleles.

442 The resulting dataset comprised 36 taxonomic references of Kp1-Kp7 (Rodrigues et al. 2019) that,
443 together with eight additional genomes of phylogroup Kp7, were considered as a reference dataset of
444 the KpSC taxa (*K. quasivariicola* reference strain KPN1705, SB6096, was excluded because it had more
445 than 30 missing alleles). Besides these 44 reference genomes, 7,154 GenBank genomes were retained,
446 resulting in a total dataset of 7,198 genomes (**Table S1; Table S2**). For some analyses, 138 genomes
447 were set aside, defined as ‘hybrids’ between phylogroups (see below), resulting in a 7,060-genome
448 dataset (**Figure S2**).

449 We estimated within-outbreak variation using previously published outbreak sets (**Table S6, Table S7**).

450

451 ***Recording sequence variation at the cgMLST gene loci***

452 Allelic variation at scgMLSTv2 loci was determined with the following strategy. First the sequence of
453 strain NTUH-K2044 was used as the reference genome, with all its alleles defined as allele 1. Then,
454 BLASTN searches (70% identity, 90% length coverage) were carried out using allele 1 as query against
455 the genomic sequences of reference genomes 18A069, 342, 01A065, 07A044, CDC4241-71 and
456 08A119, representing major lineages (phylogroups Kp2 to Kp6, including two genomes of Kp3 and
457 excluding Kp7, which was not discovered yet) of the KpSC (Blin et al. 2017). Only sequences with a
458 complete CDS (start and stop, no internal frameshift) and within a plus/minus 5% range of the
459 reference size were accepted. Alleles defined from these reference genomes and from NTUH-K2044
460 were then defined as type alleles.

461 New alleles were identified by BLASTN searches using a 90% identity threshold, 90% length coverage
462 and a word size of 30 and the above defined type alleles. The use of type alleles avoided expanding
463 the sequence space of alleles in an uncontrolled way, at the cost of losing a few highly divergent alleles,
464 which may have replaced original (vertically inherited) alleles by horizontal gene transfer (HGT) and
465 homologous recombination. As for type alleles, novel alleles were accepted only if they (i)
466 corresponded to a complete CDS (start and stop codons with no internal frameshift mutations) and (ii)
467 were within a 5% (plus/minus) of the size of the type allele size. Novel allele sequences were also

468 excluded if they came from assemblies with more than 500 contigs of size > 200 nt, as these may
469 correspond to low quality assemblies and that might contain artifactual alleles. Genome assemblies
470 based on 454 sequencing technology, which are prone to frameshifts, were also excluded for novel
471 allele definitions. No genome assemblies based on IonTorrent sequencing technology were found.

472 In order to speed the scanning process, we used the fast scan option (-e -f) of the BIGSdb autotag.pl
473 script (https://bigsdbs.readthedocs.io/en/latest/offline_tools.html). This option limits the BLASTN
474 search to a few exemplar alleles, which are used as query to find the genomic region corresponding to
475 the locus. In a second step, a direct database lookup of the region was performed to identify the exact
476 allele.

477

478 ***Definition of MLST sequence types (ST) and core genome MLST sequence types (cgST)***

479 Classical 7-gene MLST loci have been defined previously (Diancourt et al. 2005) as internal portions of
480 the seven protein-coding genes *gapA*, *infB*, *mdh*, *pgi*, *phoE*, *rpoB* and *tonB*. Novel alleles were defined
481 in the Institut Pasteur *Klebsiella* MLST and whole-genome MLST database
482 <https://bigsdbs.pasteur.fr/klebsiella/>. In 7-locus MLST, the combination of the seven allelic numbers
483 determines the isolate profile, and each unique profile is attributed a sequence type (ST) number.
484 Incomplete MLST profiles with one (or more) missing gene(s) are recorded in the isolates database but
485 in these cases, no ST number can be attributed and the profiles are therefore not defined in the
486 sequence definition database. The 7-locus MLST genes were not included in the scgMLSTv2 scheme.

487 Similar to ST identifiers used for unique 7-gene MLST allelic combinations, each distinct cgMLST profile
488 can be assigned a unique identifier; however, when using draft genomes, cgMLST data can be partly
489 incomplete due to *de novo* assembly shortcomings or missing loci. cgSTs were therefore defined only
490 for cgMLST profiles with no more than 30 uncalled alleles out of the 629 cgMLSTv2 loci. In addition,
491 we have used the --match_missing option of the define_profiles.pl script, which allows missing loci to
492 be treated as specific alleles rather than 'any' alleles. While this retains more information (because it
493 differentiates profiles that differ only by missing data at different loci), it can result in some isolates
494 genomes corresponding potentially to more than a single cgST; their equality can nevertheless be
495 deduced by the last MLST level, 'mismatch 0', as these will be grouped into the same clusters at level
496 0; this is because the clustering ignores loci with missing data in any of the profiles of a pair when
497 calculating the pairwise distance. For example, cgST1 = 0-N-1-1; cgST2 = 0-2-N-1; an isolate with profile
498 0-2-1-1 would result in cgST3 = 0-2-1-1 being created, and this genome would equate to both previous
499 cgSTs and would be labeled as cgST1; cgST2; cgST3.

500 ***Phylogenetic analyses, recombination tests and screens for virulence and resistance genes***

501 JolyTree v2.0 (Criscuolo 2019; 2020) was used to reconstruct a phylogenetic tree of the KpSC. For this,
502 first a single linkage clustering was performed to cluster cgSTs into partitions. This clustering was
503 applied on the pairwise distances between allelic profiles, defined as the number of loci with different
504 alleles, normalized by the number of loci with alleles called in both profiles. A threshold of 8
505 mismatches was defined, resulting in 2,417 clusters. One genome from each of these 2,417 clusters
506 was used as an exemplar for phylogenetic analysis (**Figure 1**).

507 A core genome multiple sequence alignment (cg-MSA) of 7,060 cgMLST profiles free of evidence for
508 inter-phylogroup ‘hybridization’ (see below) was constructed. The gene sequences were retrieved
509 based on allele number in the sequence definition database, individual gene sequences were aligned
510 with MAFFT v7.467 (missing alleles were converted into gaps), and the multiple sequence alignments
511 were concatenated. IQ-TREE v2.0.6 was used to infer a phylogenetic tree with the GTR+G
512 model (**Figure 2**).

513 Locus-by-locus recombination analyses were computed with the PHI test (Bruen, Philippe, and Bryant
514 2006) using PhiPack v1.0.

515 Kleborate v2.0.4 (Lam et al. 2021) was employed to identify acquired antimicrobial resistance and
516 virulence genes in genomic sequences, based on CARD v3.0.8 database, with identity >80% and
517 coverage >90%. Virulence score (ranges from 0 to 5) and antimicrobial resistance score (ranges from 0
518 to 3) were also derived from Kleborate. The virulence score is assigned according to the presence of
519 yersiniabactin (*ybt*), colibactin (*clb*) and aerobactin (*iuc*), as follows: 0 = none present,
520 1 = yersiniabactin only, 2 = yersiniabactin and colibactin (or colibactin only), 3 = aerobactin (without
521 yersiniabactin or colibactin), 4 = aerobactin and yersiniabactin (without colibactin), and 5 = all three
522 present. Resistance scores are calculated as follows: 0 = no ESBL (Extended-Spectrum Beta-
523 Lactamases), no carbapenemase, 1 = ESBL without carbapenemase, 2 = carbapenemase without
524 colistin resistance, 3 = carbapenemase with colistin resistance.

525

526 ***Detection of hybrid genomes***

527 Horizontal gene transfer of large portions of the genome can occur among isolates belonging to distinct
528 KpSC phylogroups (Holt et al. 2015). Additionally, MLST or scgMLST alleles may have been transferred
529 horizontally from non-KpSC members, for example *E. coli*. For the purpose of phylogeny-based
530 classification, putative hybrid genomes were excluded. To define genomes that result from large inter-

531 phylogroup recombination events, the gene-by-gene approach was used to define an original strategy,
532 outlined briefly here and more thoroughly in the supplementary appendix (Detection of hybrids): for
533 each locus, each allele was unambiguously labelled by one of the seven KpSC phylogroup of origin, if
534 possible; next, for each profile, a phylogroup homogeneity index (*i.e.*, proportion of alleles labelled by
535 the predominant phylogroup) was derived. The distributions of the phylogroup homogeneity indices
536 allowed determining hybrid genomes (**Figure S12**). Exclusion of such hybrid genomes resulted in a
537 genomic dataset of 7,060 isolates deemed as having a majority of alleles inherited from within a single
538 phylogroup. Of the 44 reference genomes, one (SB1124, of phylogroup Kp2) was defined as having a
539 hybrid origin: 414 alleles were attributed to Kp2, whereas 150 alleles originated from non-KpSC
540 species; as a result, 73.4% of SB1124 alleles were part of the majority phylogroup, which was below
541 the defined threshold of 78%. The quantification of recombination breakpoints was performed based
542 on the position of cgMLST loci on the NTUH-K2044 reference genome (NC_012731), counting the
543 number of recombination breakpoints in each successive 500 kb fragment along the reference
544 genome. Note that hybrids had typical assembly sizes (**Table S1**), whereas our simulations of
545 contaminated sequence read sets between phylogroups resulted in significantly larger assemblies (not
546 shown; available upon request).

547

548 ***Identification of genetic discontinuities in the KpSC population structure***

549 The tool MSTclust v0.21b (<https://gitlab.pasteur.fr/GIPhy/MSTclust>) was used to perform the single
550 linkage clustering of cgMLST profiles from their pairwise allelic mismatch dissimilarities, as well as to
551 assess the efficiency of the resulting profile partitioning (for details, see supplementary appendix:
552 Minimum Spanning tree-based clustering of cgMLST profiles). Briefly, for each threshold t (= 0 to 629
553 allelic mismatches), the clustering consistency was assessed using the silhouette metrics S_t (Rousseeuw
554 1987), whereas its robustness to profile subsampling biases was assessed using a dedicated metrics W_t
555 based on the adjusted Wallace coefficients (Wallace 1983; Severiano et al. 2011). Both consistency (S_t)
556 and stability (W_t) coefficients converge to 1 when the threshold t leads to a clustering that is consistent
557 with the 'natural' grouping and is robust to subsampling biases, respectively.

558 The adjusted Rand index R_t (Carrico et al. 2006; Hubert and Arabie 1985) was used to assess the global
559 concordance between single linkage clustering partitions and those induced by classifications into 7-
560 gene MLST sequence types, subspecies and species.

561

562 ***Diversity and phylogenetic compatibility indices***

563 Simpson's diversity index was computed using the www.comparingpartitions.info website (Carrico et
564 al. 2006). The clade compatibility index of STs or other groups was calculated using the ETE Python
565 library ([http://etetoolkit.org/docs/latest/tutorial/tutorial_trees.html#checking-the-monophyly-of-](http://etetoolkit.org/docs/latest/tutorial/tutorial_trees.html#checking-the-monophyly-of-attributes-within-a-tree)
566 [attributes-within-a-tree](http://etetoolkit.org/docs/latest/tutorial/tutorial_trees.html#checking-the-monophyly-of-attributes-within-a-tree)), in order to define whether their constitutive genomes formed a
567 monophyletic, paraphyletic or polyphyletic group within the recombination-purged sequence-based
568 phylogeny of the core genome. We estimated clade compatibility as the proportion of non-singleton
569 STs, sublineages or clonal groups that were monophyletic.

570

571 ***Classification of cgMLST profiles into clonal groups and sublineages***

572 The classification scheme functionality was implemented within BIGSdb v1.14.0 and relies on single
573 linkage clustering. Briefly, cgSTs were defined in the sequence definitions ('seqdef') database as
574 distinct profiles with fewer than 30 missing alleles over the scgMLST scheme, and their pairwise
575 cgMLST distance was computed as the number of distinct alleles. To account for missing data, a relative
576 threshold was used for clustering: the number of allelic mismatches was multiplied by the proportion
577 of loci for which an allele was called in both strains. Hence, in order to be grouped, the number of
578 matching alleles must exceed: (the number of loci called in both strains × (total loci - defined
579 threshold)) / total loci. cgSTs and their corresponding sublineage (SL), clonal group (CG) and other
580 levels partition identifiers, are stored in the seqdef database and are publicly available. Here,
581 classification schemes were defined in the *Klebsiella* seqdef database on top of the scgMLSTv2 scheme,
582 and host single linkage clustering group identifiers at the 10 defined cgMLST allelic mismatch
583 thresholds (see Results). For classification groups defined using 43 and 190 allelic mismatch thresholds,
584 scheme fields were defined and populated with the identifiers defined by inheritance from 7-gene
585 MLST ST identifiers (see supplementary appendix: Nomenclature inheritance algorithm).

586

587

588 ***Adaptation of the LIN code approach to cgMLST: defining cgLIN codes***

589 Vinatzer and colleagues proposed an original nomenclature method in which each genome is
590 attributed a Life Identification Number code (LIN code), based on genetic similarity with the closest
591 previously encoded member of the nomenclature (Marakeby et al. 2014; Weisberg et al. 2015;

592 Vinatzer et al. 2016; Vinatzer, Tian, and Heath 2017; Tian et al. 2020). In this proposal, the similarity
593 between genomes was based on ANI (Average Nucleotide Identity; (Konstantinidis and Tiedje 2005;
594 Goris et al. 2007)), with a set of 24 thresholds corresponding to ANI percentages of 60, 70, 75, 80, 85,
595 90, 95, 98, 98.5, 99, 99.25, 99.5, 99.75, 99.9, 99.925, 99.95, 99.975, 99.99, 99.999 and 99.9999. Here,
596 the method was adapted by replacing the ANI metric by the similarity between cgMLST profiles,
597 defined as the proportion of loci with identical alleles normalized by the number of loci with alleles
598 called in both profiles. These codes, which we refer to as cgLIN codes, are composed of a set of p
599 positions, each corresponding to a pairwise genome similarity threshold s_p . These similarity thresholds
600 are sorted in ascending order (*i.e.*, $s_p < s_{p+1}$), the first positions of the code (on the left side) thus
601 corresponding to low levels of similarity. Following the initial proposal, the codes are assigned as
602 follows (**Figure S13**): (step 1) the code is initialized with the first strain being assigned the value "0" at
603 all positions; (step 2) the encoding rule for a new genome i is based on the closest genome j already
604 encoded as follows, from the similarity $s_{ij} \in] s_{p-1}, s_p]$:

- 605 i) identical to code j up to and including position $p - 1$;
- 606 ii) for the position p : maximum value observed at this position (among the subset of codes
607 sharing the same prefix at the position $p - 1$) incremented by 1;
- 608 iii) "0" to all downstream positions, from $p + 1$ included.

609 For each genome to be encoded, step 2 is repeated.

610

611 A set of 10 cgMLST thresholds were defined as follows: first, four thresholds were chosen above the
612 similarity values peak observed between *Klebsiella* species ($s_p = 1 - 610 / 629 = 0.03$), subspecies ($s_p =$
613 $1 - 585 / 629 = 0.07$), main sublineages ($s_p = 1 - 190 / 629 = 0.70$) and clonal groups ($s_p = 1 - 43 / 629 =$
614 0.93). Second, we included six thresholds deemed useful for epidemiological studies, corresponding to
615 10, 7, 4, 2, 1 and 0 allelic mismatches.

616 This encoding system conveys phylogenetic information, as two genomes with identical prefixes in
617 their respective cgLIN codes can be understood as being similar, to an extent determined by the length
618 of their common prefix. Isolates having cgMLST profiles with 100% identity (no mismatch at loci called
619 in both genomes) will have exactly the same cgLIN code. For example, cgLIN codes
620 0_0_22_12_0_1_0_0_0_0 and 4_0_3_0_0_0_0_0_0_0 would denote two strains belonging to distinct
621 species (as they differ by their first number in the code). cgLIN codes 0_0_105_6_0_0_75_1_1_0 and
622 0_0_105_6_0_0_75_1_0_0 correspond to strains from Kp1 (prefix 0_0) that differ by only 2 loci; they
623 are identical up to the second bin, corresponding to 2 locus mismatches (**Figure S13**); note that 0 and

624 1 mismatches are both included in the last bin: genomes have an identical identifier when having 0
625 difference, and a different identifier when having 1 mismatch (**Figure S14**).

626 The impact of genome input order on the number of cgLIN code partitions at a given threshold was
627 defined using the 7,060 high-quality, non-hybrid cgMLST profiles, which were encoded 500 times with
628 random input orders (see details in the supplementary appendix: Impact of strains input order on LIN
629 codes, and use of Prim's algorithm).

630 The scripts for cgLIN code database creation were made available via GitLab BEBP
631 (<https://gitlab.pasteur.fr/BEBP/LINcoding>).

632 **Acknowledgements**

633 This work builds partly on the classical 7-gene MLST database, which has been curated by Virginie
634 Passet, Radek Izdebski, Carla Rodrigues and Federica Palma. We thank Adrien Le Meur for help with
635 the genome dataset constitution and initial analyses of intra-outbreak variation.

636 **Funding**

637 MH was supported financially by the PhD grant “Codes4strains” from the European Joint Programme
638 One Health, which has received funding from the European Union’s Horizon 2020 Research and
639 Innovation Programme under Grant Agreement No. 773830. This work was supported financially by
640 the French Government’s Investissement d’Avenir program Laboratoire d’Excellence “Integrative
641 Biology of Emerging Infectious Diseases” (ANR-10-LABX-62-IBEID) and by an award from the Bill and
642 Melinda Gates Foundation (INV-025280). BIGSdb development is funded by a Wellcome Trust
643 Biomedical Resource grant (218205/Z/19/Z). This work used the computational and storage services
644 provided by the IT department at Institut Pasteur.

645 **Authors license statement**

646 This research was funded, in whole or in part, by Institut Pasteur and by European Union’s Horizon
647 2020 research and innovation programme. For the purpose of open access, the authors have applied
648 a CC-BY public copyright license to any Author Manuscript version arising from this submission.

649 **Declaration of interest statement:** The authors declare no conflict of interest.

650 **Ethical approval statement:** Not relevant.

651 **Author contributions**

652 S.B. designed and coordinated the study. M.H. performed the genomic analyses, and cgLIN code
653 developments and implementations. J.G. designed the novel version of the cgMLST scheme. A.C.
654 developed the MSTclust tool (and associated metrics), and supervised cgLIN developments and
655 phylogenetic analyses. K.A.J. and M.C.J.M. designed and developed the BIGSdb platform, with help
656 from S. Bridel for the integration of cgLIN code functionality. S.B. and M.H. wrote the initial version of
657 the manuscript, with input from A.C. All authors provided input to the manuscript and reviewed the
658 final version.

659 **Figure legends**

660 **Figure 1. Genome-based phylogenetic tree of the *K. pneumoniae* species complex.**

661 The whole-genome distance-based tree was inferred using JolyTree. JolyTree uses *mash* to decompose
662 each genome into a sketch of k-mers and to quickly estimate the p-distance between each pair of
663 genomes; after transforming every p-distance into a pairwise evolutionary distance, a phylogenetic
664 tree is inferred using FastME. The seven phylogroups are indicated. Red dots correspond to strains
665 defined as inter-phylogroup hybrids. Scale bar, 0.01 nucleotide substitutions per site.

666 **Figure 2. Phylogenetic structure within phylogroup Kp1 (*K. pneumoniae sensu stricto*).**

667 The circular tree was obtained using IQ-TREE based on the concatenation of the genes of the
668 scgMLSTv2 scheme; 1,600 isolates are included (see Methods). Labels on the external first circle
669 represent 7-gene MLST ST identifiers (each alternation corresponds to a different ST and only ST with
670 more than 20 strains are labelled). The second and third circles (light green and blue, respectively)
671 show the alternation of clonal groups (CG) and sublineages (SL), respectively, labelling only groups with
672 more than 20 isolates. Full correspondence between ST, SL and CG identifiers is given in the
673 supplementary appendix.

674 **Figure 3. Distribution of pairwise cgMLST distances, clustering properties and phylogenetic 675 congruence.**

676 Values are plotted for the 7,060 genomes dataset. Threshold values (t) are shown on the X-axis,
677 corresponding to allelic profile mismatch values up to 629 (or 100%). Grey histograms: distribution of
678 pairwise allelic mismatches. The circles correspond to the different modes of distribution. The curves
679 represent the consistency (silhouette) and stability coefficients S_t (blue) and W_t (green), respectively,
680 obtained with each threshold t ; the corresponding scale is on the left Y-axis. The dotted vertical red
681 lines at $t = 43/629$, $190/629$, $585/629$ and $610/629$ represent the thresholds up to which pairs of
682 genomes belong to the same clonal groups, sublineages, phylogroups and species, respectively.

683 **Figure 4. Concordance of sublineage, clonal group and 7-gene MLST classifications.**

684 Alluvial diagram obtained using RAWGraphs (Mauri *et al.* 2017: doi.org/10.1145/3125571.3125585)
685 showing the correspondence between sequence types (ST; 7 genes identity), clonal groups (CG; 43
686 allelic mismatches threshold) and sublineages (SL; 190 allelic mismatches threshold). Colors are
687 arbitrarily attributed by the software for readability.

688 **Figure 5. Phylogenetic relationships are reflected in cgLIN code prefixes.**

689 Left: the prefix tree generated from cgLIN codes; Right: phylogenetic relationships derived using IQ-
690 TREE from the cgMLST gene sequences from the reference strains. The cgLIN codes are also shown.
691 The values indicated on top of the prefix tree correspond to the cgMLST similarity percentage of the
692 corresponding cgLIN code bin.

693 **Figure 6. cgLIN code prefixes, and virulence and antimicrobial resistance scores of some sublineages**
694 **and their clonal groups.**

695 Left (green) panel: LIN prefixes of selected sublineages (SL) and clonal groups (CG). Right: heatmaps of
696 virulence and resistance scores of clonal groups, and the number of genomes in each group. For each
697 genome, the virulence score derived from Kleborate has a value from 0 to 5; the value in the cells
698 corresponds to the percentage of strains in the group with that virulence score (similar to a heat map).
699 The principle is the same for the resistance score, but it varies from 0 to 3.

700

701 **Tables**

702 **Table 1. Genome dataset phylogroup breakdown, quality assessment and diversity.**

703 **References**

- 704 Achtman, Mark, John Wain, François-Xavier Weill, Satheesh Nair, Zheming Zhou, Vartul Sangal, Mary G.
 705 Krauland, et al. 2012. 'Multilocus Sequence Typing as a Replacement for Serotyping in
 706 *Salmonella Enterica*'. *PLoS Pathog* 8 (6): e1002776.
 707 <https://doi.org/10.1371/journal.ppat.1002776>.
- 708 Barker, Dillon OR, João A. Carriço, Peter Kruczkiewicz, Federica Palma, Mirko Rossi, and Eduardo N.
 709 Taboada. 2018. 'Rapid Identification of Stable Clusters in Bacterial Populations Using the
 710 Adjusted Wallace Coefficient'. *BioRxiv*, April, 299347. <https://doi.org/10.1101/299347>.
- 711 Belkum, A. van, P. T. Tassios, L. Dijkshoorn, S. Haeggman, B. Cookson, N. K. Fry, V. Fussing, et al. 2007.
 712 'Guidelines for the Validation and Application of Typing Methods for Use in Bacterial
 713 Epidemiology'. *Clin Microbiol Infect* 13 Suppl 3 (October): 1–46.
- 714 Bialek-Davenet, S., A. Criscuolo, F. Ailloud, V. Passet, L. Jones, A. S. Delannoy-Vieillard, B. Garin, et al.
 715 2014. 'Genomic Definition of Hypervirulent and Multidrug-Resistant *Klebsiella Pneumoniae*
 716 Clonal Groups'. *Emerg Infect Dis* 20 (11): 1812–20. <https://doi.org/10.3201/eid2011.140206>.
- 717 Blin, C., V. Passet, M. Touchon, E. P. C. Rocha, and S. Brisse. 2017. 'Metabolic Diversity of the Emerging
 718 Pathogenic Lineages of *Klebsiella Pneumoniae*'. *Environ Microbiol* 19 (5): 1881–98.
 719 <https://doi.org/10.1111/1462-2920.13689>.
- 720 Bowers, Jolene R., Brandon Kitchel, Elizabeth M. Driebe, Duncan R. MacCannell, Chandler Roe, Darrin
 721 Lemmer, Tom de Man, et al. 2015. 'Genomic Analysis of the Emergence and Rapid Global
 722 Dissemination of the Clonal Group 258 *Klebsiella Pneumoniae* Pandemic'. *PLoS One* 10 (7):
 723 e0133727. <https://doi.org/10.1371/journal.pone.0133727>.
- 724 Brisse, S., F. Grimont, and P.A.D. Grimont. 2006. 'The Genus *Klebsiella*'. In *The Prokaryotes A Handbook*
 725 *on the Biology of Bacteria*, 3rd edition, 159–96. New York: Springer.
- 726 Brisse, S., and J. Verhoef. 2001. 'Phylogenetic Diversity of *Klebsiella Pneumoniae* and *Klebsiella Oxytoca*
 727 Clinical Isolates Revealed by Randomly Amplified Polymorphic DNA, *GyrA* and *ParC* Genes
 728 Sequencing and Automated Ribotyping'. *Int J Syst Evol Microbiol* 51 (Pt 3): 915–24.
- 729 Bruen, Trevor C., Hervé Philippe, and David Bryant. 2006. 'A Simple and Robust Statistical Test for
 730 Detecting the Presence of Recombination'. *Genetics* 172 (4): 2665–81.
 731 <https://doi.org/10.1534/genetics.105.048975>.
- 732 Carrico, J. A., C. Silva-Costa, J. Melo-Cristino, F. R. Pinto, H. de Lencastre, J. S. Almeida, and M. Ramirez.
 733 2006. 'Illustration of a Common Framework for Relating Multiple Typing Methods by
 734 Application to Macrolide-Resistant *Streptococcus Pyogenes*'. *J Clin Microbiol* 44 (7): 2524–32.
 735 <https://doi.org/10.1128/JCM.02536-05>.
- 736 Chen, Liang, Barun Mathema, Johann D. D. Pitout, Frank R. DeLeo, and Barry N. Kreiswirth. 2014.
 737 'Epidemic *Klebsiella Pneumoniae* ST258 Is a Hybrid Strain'. *MBio* 5 (3): e01355-01314.
 738 <https://doi.org/10.1128/mBio.01355-14>.
- 739 Cowan, S. T. 1965. 'PRINCIPLES AND PRACTICE OF BACTERIAL TAXONOMY--A FORWARD LOOK'. *Journal*
 740 *of General Microbiology* 39 (April): 143–53. <https://doi.org/10.1099/00221287-39-1-143>.
- 741 Criscuolo, Alexis. 2019. 'A Fast Alignment-Free Bioinformatics Procedure to Infer Accurate Distance-
 742 Based Phylogenetic Trees from Genome Assemblies'. *Research Ideas and Outcomes* 5 (June):
 743 e36178. <https://doi.org/10.3897/rio.5.e36178>.
- 744 ———. 2020. 'On the Transformation of MinHash-Based Uncorrected Distances into Proper
 745 Evolutionary Distances for Phylogenetic Inference'. *F1000Research* 9 (November): 1309.
 746 <https://doi.org/10.12688/f1000research.26930.1>.
- 747 Dallman, Timothy, Philip Ashton, Ulf Schafer, Aleksey Jironkin, Anais Painset, Sharif Shaaban, Hassan
 748 Hartman, et al. 2018. 'SnapperDB: A Database Solution for Routine Sequencing Analysis of
 749 Bacterial Isolates'. *Bioinformatics (Oxford, England)* 34 (17): 3028–29.
 750 <https://doi.org/10.1093/bioinformatics/bty212>.

751 Diancourt, L., V. Passet, J. Verhoef, P. A. Grimont, and S. Brisse. 2005. 'Multilocus Sequence Typing of
752 Klebsiella Pneumoniae Nosocomial Isolates'. *J Clin Microbiol* 43 (8): 4178–82.

753 Feil, E. J. 2004. 'Small Change: Keeping Pace with Microevolution'. *Nat. Rev. Microbiol.* 2 (6): 483–95.

754 Fevre, C., V. Passet, F. X. Weill, P. A. Grimont, and S. Brisse. 2005. 'Variants of the Klebsiella
755 Pneumoniae OKP Chromosomal Beta-Lactamase Are Divided into Two Main Groups, OKP-A
756 and OKP-B'. *Antimicrob Agents Chemother* 49 (12): 5149–52.

757 Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje. 2007. 'DNA-
758 DNA Hybridization Values and Their Relationship to Whole-Genome Sequence Similarities'. *Int
759 J Syst Evol Microbiol* 57 (Pt 1): 81–91.

760 Hacker, J., and J. B. Kaper. 2000. 'Pathogenicity Islands and the Evolution of Microbes'. *Annu Rev
761 Microbiol* 54: 641–79.

762 Holt, K. E., H. Wertheim, R. N. Zadoks, S. Baker, C. A. Whitehouse, D. Dance, A. Jenney, et al. 2015.
763 'Genomic Analysis of Diversity, Population Structure, Virulence, and Antimicrobial Resistance
764 in Klebsiella Pneumoniae, an Urgent Threat to Public Health'. *Proc Natl Acad Sci U S A* 112 (27):
765 E3574-81. <https://doi.org/10.1073/pnas.1501049112>.

766 Hubert, Lawrence, and Phipps Arabie. 1985. 'Comparing Partitions'. *Journal of Classification* 2 (1): 193–
767 218. <https://doi.org/10.1007/BF01908075>.

768 Jolley, K. A., and M. C. Maiden. 2010. 'BIGSdb: Scalable Analysis of Bacterial Genome Variation at the
769 Population Level'. *BMC Bioinformatics* 11: 595. <https://doi.org/10.1186/1471-2105-11-595>.

770 Jolley, Keith A., James E. Bray, and Martin C. J. Maiden. 2018. 'Open-Access Bacterial Population
771 Genomics: BIGSdb Software, the PubMLST.Org Website and Their Applications'. *Wellcome
772 Open Research* 3: 124. <https://doi.org/10.12688/wellcomeopenres.14826.1>.

773 Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. 'Genomic Insights That Advance the
774 Species Definition for Prokaryotes'. *Proceedings of the National Academy of Sciences of the
775 United States of America* 102 (7): 2567–72. <https://doi.org/10.1073/pnas.0409727102>.

776 Lam, Margaret M. C., Ryan R. Wick, Stephen C. Watts, Louise T. Cerdeira, Kelly L. Wyres, and Kathryn
777 E. Holt. 2021. 'A Genomic Surveillance Framework and Genotyping Tool for Klebsiella
778 Pneumoniae and Its Related Species Complex'. *Nature Communications* 12 (1): 4188.
779 <https://doi.org/10.1038/s41467-021-24448-3>.

780 Lam, Margaret M. C., Kelly L. Wyres, Sebastian Duchêne, Ryan R. Wick, Louise M. Judd, Yunn-Hwen
781 Gan, Chu-Han Hoh, et al. 2018. 'Population Genomics of Hypervirulent Klebsiella Pneumoniae
782 Clonal-Group 23 Reveals Early Emergence and Rapid Global Dissemination'. *Nature
783 Communications* 9 (1): 2703. <https://doi.org/10.1038/s41467-018-05114-7>.

784 Lan, R., and P. R. Reeves. 2001. 'When Does a Clone Deserve a Name? A Perspective on Bacterial
785 Species Based on Population Genetics'. *Trends Microbiol* 9: 419–24.

786 Long, S. Wesley, Sarah E. Linson, Matthew Ojeda Saavedra, Concepcion Cantu, James J. Davis, Thomas
787 Brettin, and Randall J. Olsen. 2017. 'Whole-Genome Sequencing of a Human Clinical Isolate of
788 the Novel Species Klebsiella Quasivariicola Sp. Nov'. *Genome Announcements* 5 (42).
789 <https://doi.org/10.1128/genomeA.01057-17>.

790 Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, et al. 1998. 'Multilocus
791 Sequence Typing: A Portable Approach to the Identification of Clones within Populations of
792 Pathogenic Microorganisms'. *Proc. Natl. Acad. Sci. U. S. A.* 95 (6): 3140–45.

793 Maiden, M. C., M. J. van Rensburg, J. E. Bray, S. G. Earle, S. A. Ford, K. A. Jolley, and N. D. McCarthy.
794 2013. 'MLST Revisited: The Gene-by-Gene Approach to Bacterial Genomics'. *Nat Rev Microbiol*
795 11 (10): 728–36. <https://doi.org/10.1038/nrmicro3093>.

796 Marakeby, Haitham, Eman Badr, Hanaa Torkey, Yuhyun Song, Scotland Leman, Caroline L. Monteil,
797 Lenwood S. Heath, and Boris A. Vinatzer. 2014. 'A System to Automatically Classify and Name
798 Any Individual Genome-Sequenced Organism Independently of Current Biological
799 Classification and Nomenclature'. *PloS One* 9 (2): e89142.
800 <https://doi.org/10.1371/journal.pone.0089142>.

801 Miro, Elisenda, John W. A. Rossen, Monika A. Chlebowicz, Dag Harmsen, Sylvain Brisse, Virginie Passet,
802 Ferran Navarro, Alex W. Friedrich, and S. García-Cobos. 2020. 'Core/Whole Genome Multilocus
803 Sequence Typing and Core Genome SNP-Based Typing of OXA-48-Producing *Klebsiella*
804 *Pneumoniae* Clinical Isolates From Spain'. *Frontiers in Microbiology* 10: 2961.
805 <https://doi.org/10.3389/fmicb.2019.02961>.

806 Moura, Alexandra, Alexis Criscuolo, Hannes Pouseele, Mylène M. Maury, Alexandre Leclercq, Cheryl
807 Tarr, Jonas T. Björkman, et al. 2016. 'Whole Genome-Based Population Biology and
808 Epidemiological Surveillance of *Listeria Monocytogenes*'. *Nature Microbiology* 2 (October):
809 16185. <https://doi.org/10.1038/nmicrobiol.2016.185>.

810 Prim, R. C. 1957. 'Shortest Connection Networks And Some Generalizations'. *Bell System Technical*
811 *Journal* 36 (6): 1389–1401. <https://doi.org/10.1002/j.1538-7305.1957.tb01515.x>.

812 Rambaut, Andrew, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis,
813 Louis du Plessis, and Oliver G. Pybus. 2020. 'A Dynamic Nomenclature Proposal for SARS-CoV-
814 2 Lineages to Assist Genomic Epidemiology'. *Nature Microbiology* 5 (11): 1403–7.
815 <https://doi.org/10.1038/s41564-020-0770-5>.

816 Rodrigues, Carla, Virginie Passet, Andrianiaina Rakotondrasoa, Thierno Abdoulaye Diallo, Alexis
817 Criscuolo, and Sylvain Brisse. 2019. 'Description of *Klebsiella Africanensis* Sp. Nov., *Klebsiella*
818 *Variicola* Subsp. *Tropicalensis* Subsp. Nov. and *Klebsiella Variicola* Subsp. *Variicola* Subsp. Nov.'.
819 *Research in Microbiology*, February. <https://doi.org/10.1016/j.resmic.2019.02.003>.

820 Rousseeuw, Peter J. 1987. 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster
821 Analysis'. *Journal of Computational and Applied Mathematics* 20 (November): 53–65.
822 [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).

823 Selander, R. K., and B. R. Levin. 1980. 'Genetic Diversity and Structure in *Escherichia Coli* Populations'.
824 *Science* 210 (4469): 545–47.

825 Severiano, Ana, Francisco R. Pinto, Mário Ramirez, and João A. Carriço. 2011. 'Adjusted Wallace
826 Coefficient as a Measure of Congruence between Typing Methods'. *Journal of Clinical*
827 *Microbiology* 49 (11): 3997–4000. <https://doi.org/10.1128/JCM.00624-11>.

828 Sneath, P. H. A. 1992. *International Code of Nomenclature of Bacteria*. 1990 revision. Washington, D.C.:
829 American Society for Microbiology.

830 Struelens, M. J., Y. De Gheldre, and A. Deplano. 1998. 'Comparative and Library Epidemiological Typing
831 Systems: Outbreak Investigations versus Surveillance Systems'. *Infect Control Hosp Epidemiol*
832 19 (8): 565–69.

833 Struve, Carsten, Chandler C. Roe, Marc Stegger, Steen G. Stahlhut, Dennis S. Hansen, David M.
834 Engelthaler, Paal S. Andersen, Elizabeth M. Driebe, Paul Keim, and Karen A. Krogfelt. 2015.
835 'Mapping the Evolution of Hypervirulent *Klebsiella Pneumoniae*'. *MBio* 6 (4): e00630.
836 <https://doi.org/10.1128/mBio.00630-15>.

837 Tian, Long, Chengjie Huang, Reza Mazloom, Lenwood S Heath, and Boris A Vinatzer. 2020. 'LINbase: A
838 Web Server for Genome-Based Identification of Prokaryotes as Members of Crowdsourced
839 Taxa'. *Nucleic Acids Research* 48 (W1): W529–37. <https://doi.org/10.1093/nar/gkaa190>.

840 Turner, K. M., and E. J. Feil. 2007. 'The Secret Life of the Multilocus Sequence Type'. *Int J Antimicrob*
841 *Agents* 29 (2): 129–35.

842 Vinatzer, Boris A., Long Tian, and Lenwood S. Heath. 2017. 'A Proposal for a Portal to Make Earth's
843 Microbial Diversity Easily Accessible and Searchable'. *Antonie van Leeuwenhoek* 110 (10):
844 1271–79. <https://doi.org/10.1007/s10482-017-0849-z>.

845 Vinatzer, Boris A., Alexandra J. Weisberg, Caroline L. Monteil, Haitham A. Elmarakeby, Samuel K.
846 Sheppard, and Lenwood S. Heath. 2016. 'A Proposal for a Genome Similarity-Based Taxonomy
847 for Plant-Pathogenic Bacteria That Is Sufficiently Precise to Reflect Phylogeny, Host Range, and
848 Outbreak Affiliation Applied to *Pseudomonas Syringae* Sensu Lato as a Proof of Concept'.
849 *Phytopathology*® 107 (1): 18–28. <https://doi.org/10.1094/PHYTO-07-16-0252-R>.

850 Vos, Michiel, and Xavier Didelot. 2009. 'A Comparison of Homologous Recombination Rates in Bacteria
851 and Archaea'. *The ISME Journal* 3 (2): 199–208. <https://doi.org/10.1038/ismej.2008.93>.

- 852 Wallace, David L. 1983. 'A Method for Comparing Two Hierarchical Clusterings: Comment'. *Journal of*
853 *the American Statistical Association* 78 (383): 569–76. <https://doi.org/10.2307/2288118>.
- 854 Weisberg, Alexandra J., Haitham A. Elmarakeby, Lenwood S. Heath, and Boris A. Vinatzer. 2015.
855 'Similarity-Based Codes Sequentially Assigned to Ebolavirus Genomes Are Informative of
856 Species Membership, Associated Outbreaks, and Transmission Chains'. *Open Forum Infectious*
857 *Diseases* 2 (ofv024). <https://doi.org/10.1093/ofid/ofv024>.
- 858 Wyres, Kelly L., Claire Gorrie, David J. Edwards, Heiman F. L. Wertheim, Li Yang Hsu, Nguyen Van Kinh,
859 Ruth Zadoks, Stephen Baker, and Kathryn E. Holt. 2015. 'Extensive Capsule Locus Variation and
860 Large-Scale Genomic Recombination within the *Klebsiella Pneumoniae* Clonal Group 258'.
861 *Genome Biology and Evolution* 7 (5): 1267–79. <https://doi.org/10.1093/gbe/evv062>.
- 862 Wyres, Kelly L., Margaret M. C. Lam, and Kathryn E. Holt. 2020. 'Population Genomics of *Klebsiella*
863 *Pneumoniae*'. *Nature Reviews. Microbiology* 18 (6): 344–59. [https://doi.org/10.1038/s41579-](https://doi.org/10.1038/s41579-019-0315-1)
864 [019-0315-1](https://doi.org/10.1038/s41579-019-0315-1).
- 865 Wyres, Kelly L., To N. T. Nguyen, Margaret M. C. Lam, Louise M. Judd, Nguyen van Vinh Chau, David A.
866 B. Dance, Margaret Ip, et al. 2020. 'Genomic Surveillance for Hypervirulence and Multi-Drug
867 Resistance in Invasive *Klebsiella Pneumoniae* from South and Southeast Asia'. *Genome*
868 *Medicine* 12 (1): 11. <https://doi.org/10.1186/s13073-019-0706-y>.
- 869 Zhou, Zhemin, Jane Charlesworth, and Mark Achtman. 2021. 'HierCC: A Multi-Level Clustering Scheme
870 for Population Assignments Based on Core Genome MLST'. *Bioinformatics (Oxford, England)*,
871 April, btab234. <https://doi.org/10.1093/bioinformatics/btab234>.
- 872

Figure 1

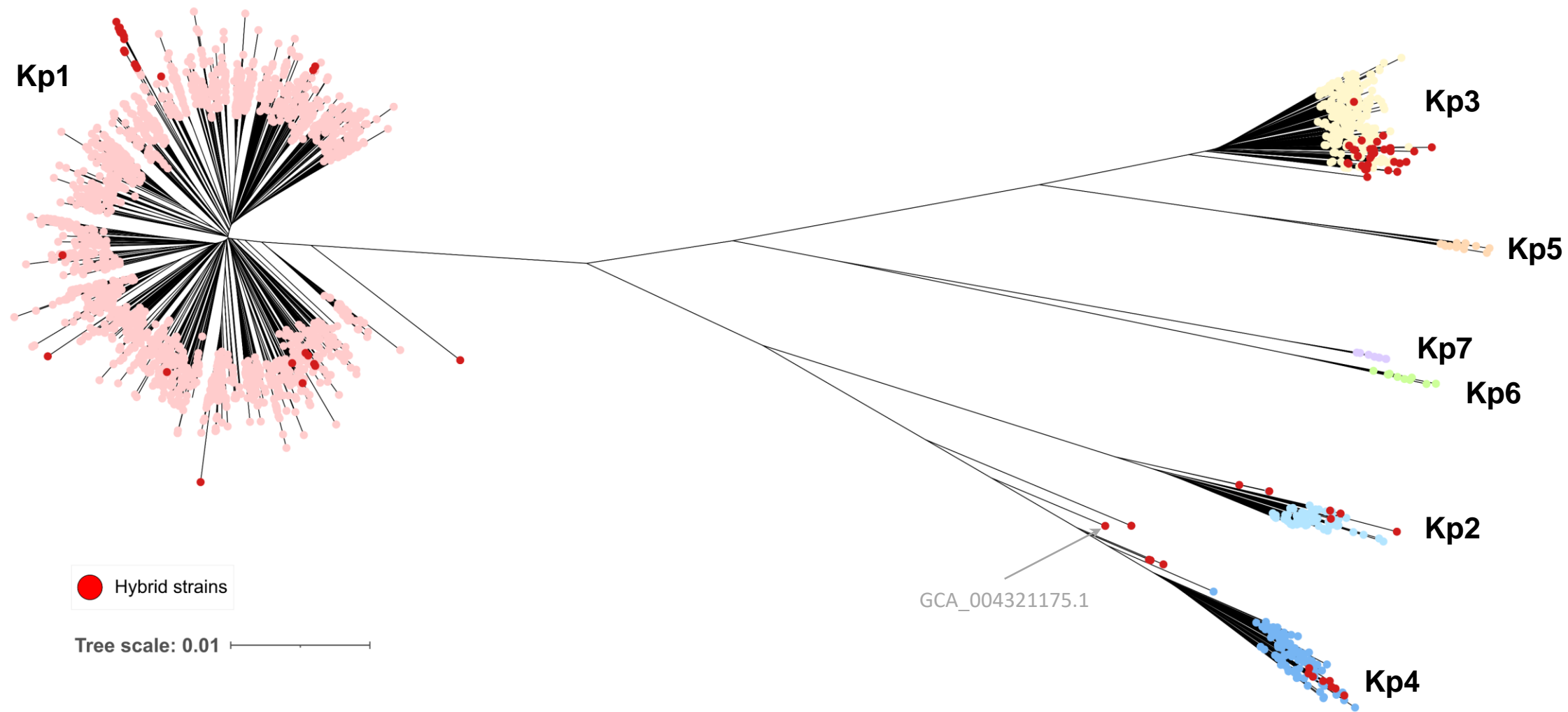


Figure 3

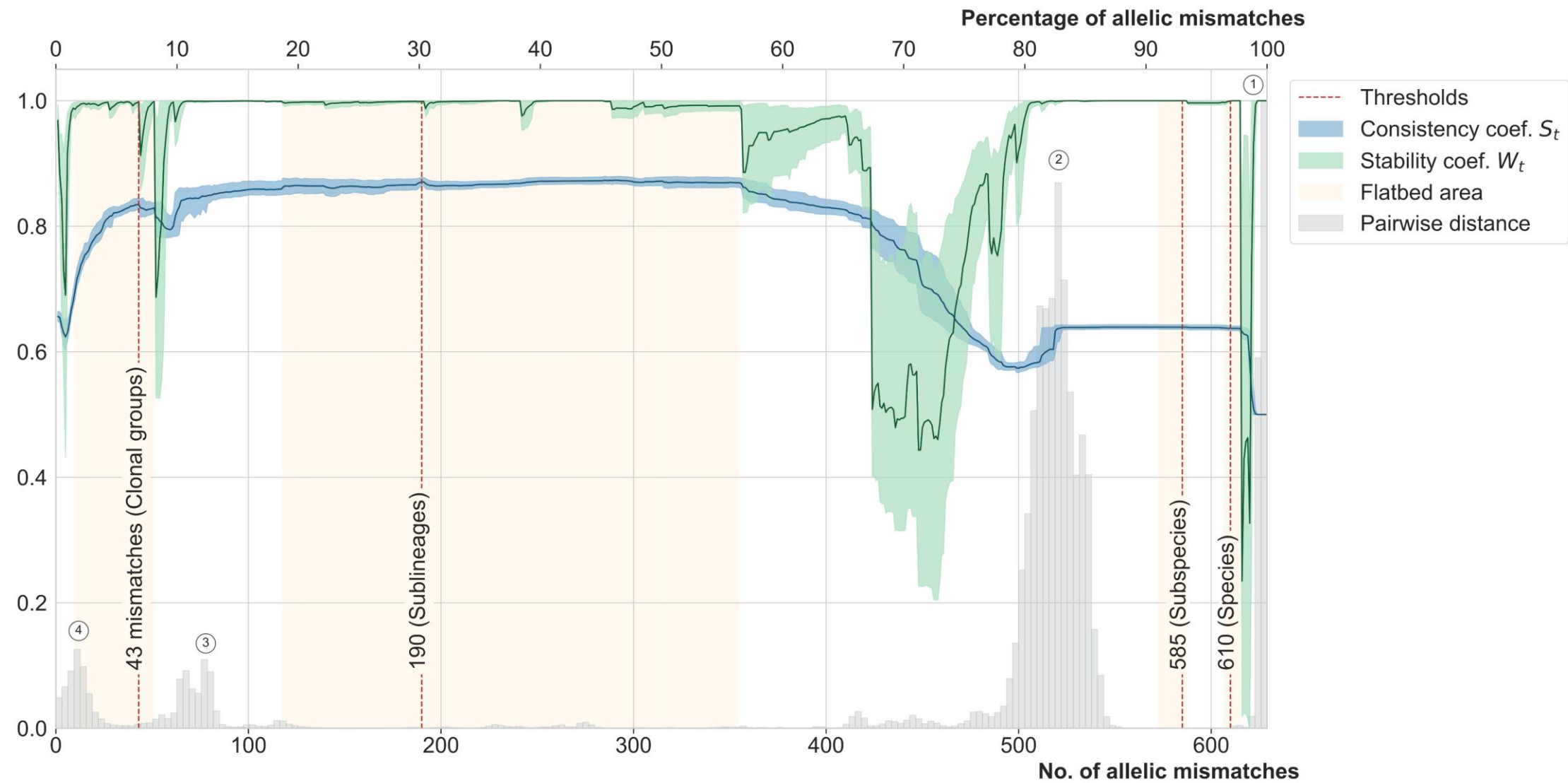
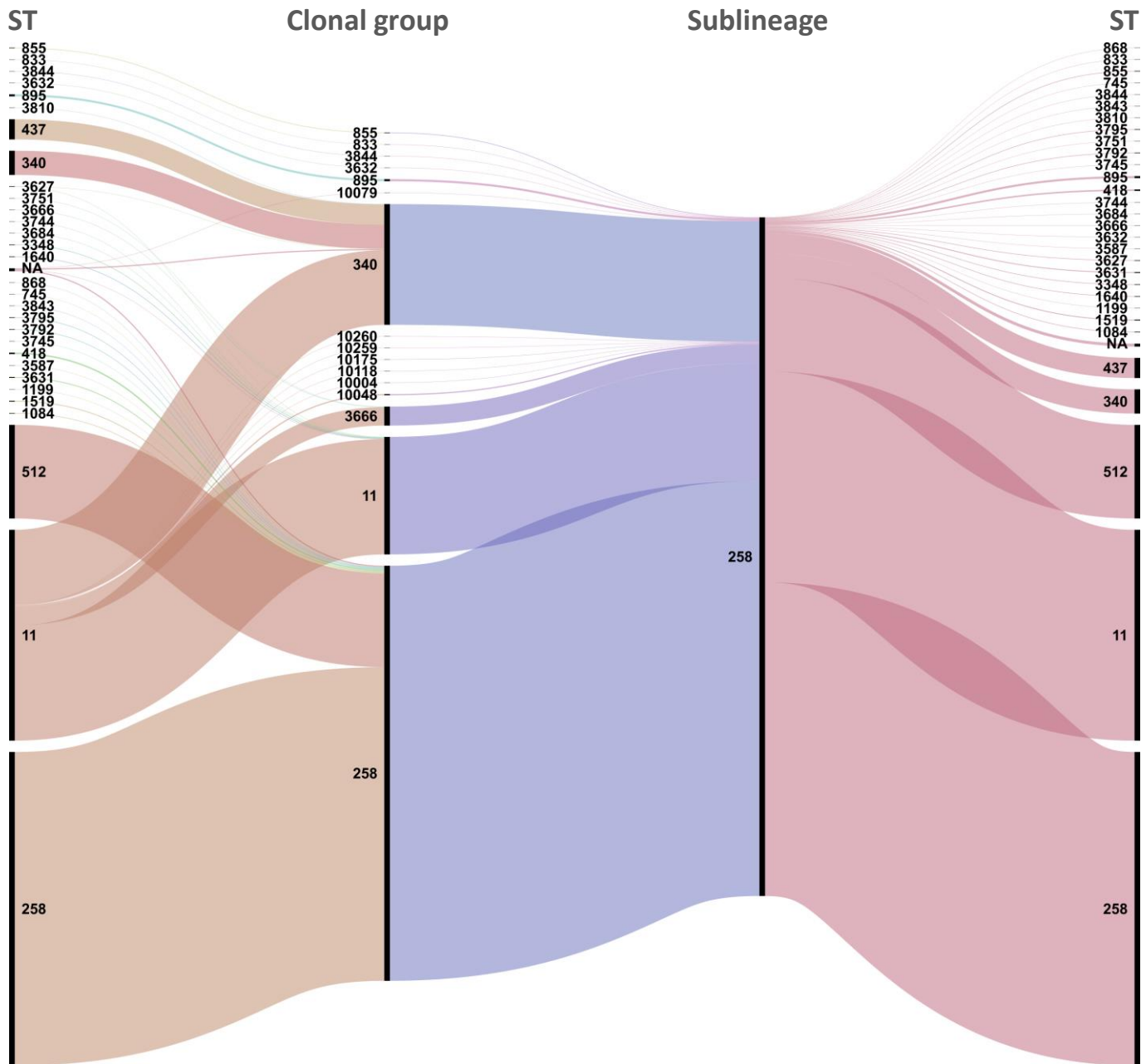


Figure 4

A. ST258 and related genomes



B. ST23 and related genomes

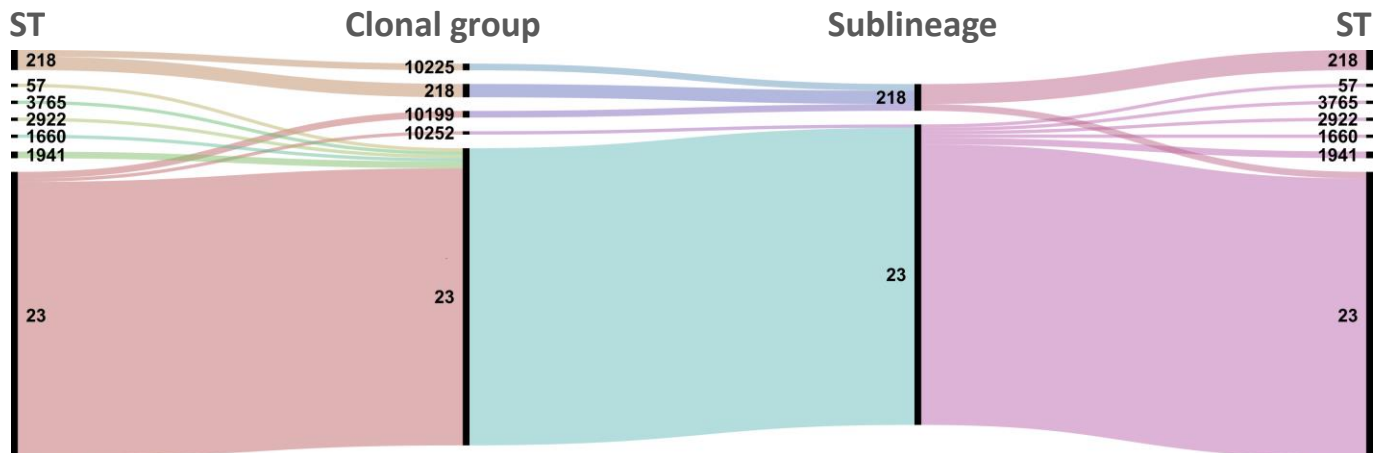


Figure 5

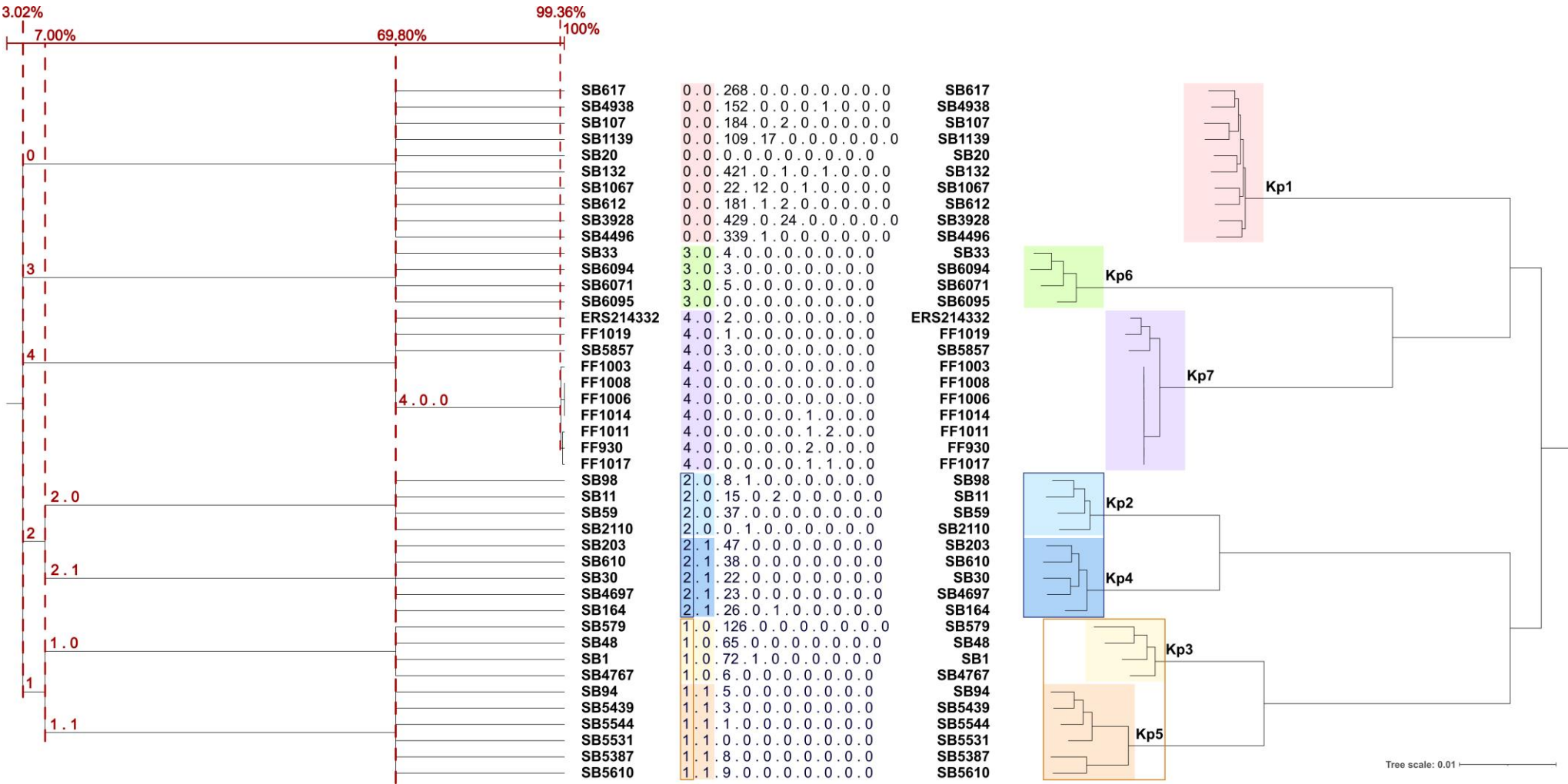


Figure 6

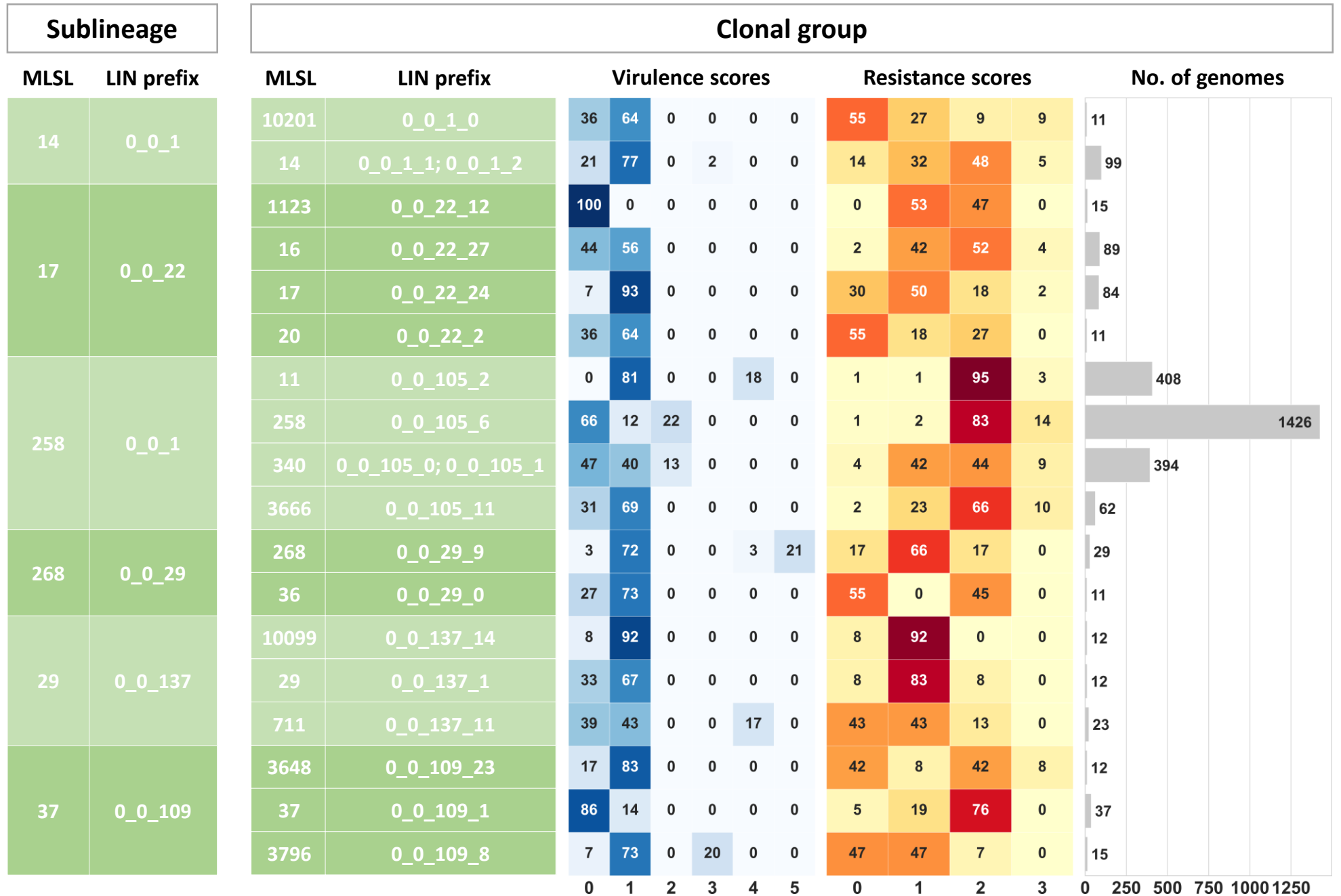


Table 1. Genome dataset phylogroup breakdown, quality assessment and diversity

Taxonomic Designation	Phylogroup	Initial No. Genomes	No. QC-filtered	No. after applying filters	No. of hybrids	No. of non-hybrid genomes	No. Called alleles (mean, std)	No. of sequence types (ST)
<i>K. pneumoniae</i> subsp. <i>pneumoniae</i>	Kp1	6737	218 (3.2%)	6519 (90.6%)	43 (0.7%)	6476 (91.7%)	624.6 (3.6)	705
<i>K. quasipneumoniae</i> subsp. <i>quasipneumoniae</i>	Kp2	115	1 (0.9%)	114 (1.6%)	8 (7.0%)	106 (1.5%)	604.0 (2.5)	49
<i>K. variicola</i> subsp. <i>variicola</i>	Kp3	309	8 (2.6%)	301 (4.2%)	37 (12.3%)	264 (3.7%)	615.4 (3.3)	149
<i>K. quasipneumoniae</i> subsp. <i>similipneumoniae</i>	Kp4	230	6 (2.6%)	224 (3.1%)	50 (22.3%)	174 (2.5%)	607.4 (1.9)	64
<i>K. variicola</i> subsp. <i>tropica</i>	Kp5	19	0 (0.0%)	19 (0.3%)	0 (0.0%)	19 (0.3%)	611.7 (1.8)	13
<i>K. quasivariicola</i>	Kp6	13	2 (15.4%)	11 (0.2%)	0 (0.0%)	11 (0.2%)	602.9 (1.8)	8
<i>K. africana</i>	Kp7	10	0 (0.0%)	10 (0.1%)	0 (0.0%)	10 (0.1%)	606.3 (1.6)	4
Total		7433	235	7198	138	7060	mean 610.3 (2.36)	992