Genomic surveillance framework and global population structure for *Klebsiella pneumoniae*

Margaret M. C. Lam[1*], Ryan R. Wick[1], Stephen C. Watts[2], Louise T. Cerdeira[1], Kelly L. Wyres[1], Kathryn E. Holt[1,3]

[1]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia. [2]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia. [3]London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK.

*Corresponding author: margaret.lam@monash.edu

ABSTRACT

K. pneumoniae is a leading cause of antimicrobial-resistant (AMR) healthcare-associated infections, neonatal sepsis and community-acquired liver abscess, and is associated with chronic intestinal diseases. Its diversity and complex population structure pose challenges for analysis and interpretation of *K. pneumoniae* genome data. Here we introduce Kleborate, a tool for analysing genomes of *K. pneumoniae* and its associated species complex, which consolidates interrogation of key features of proven clinical importance. Kleborate provides a framework to support genomic surveillance and epidemiology in research, clinical and public health settings. To demonstrate its utility we apply Kleborate to analyse publicly available *Klebsiella* genomes, including clinical isolates from a pan-European study of carbapenemase-producing *Klebsiella*, highlighting global trends in AMR and virulence as examples of what could be achieved by applying this genomic framework within more systematic genomic surveillance efforts. We also demonstrate the application of Kleborate to detect and type *K. pneumoniae* from gut metagenomes.

53

54 **TEXT**

55

56 *Klebsiella pneumoniae* bacteria commonly colonize the mammalian gut, but are also
57 recognized as a major public health threat due to their ability to cause severe
58 infections in healthcare settings and their association with antimicrobial resistance
59 (AMR)[1,2]. Reports of *K. pneumoniae* gut colonization frequencies vary by country
60 and demographics, but prevalence rates as high as 87% have been reported[3–6]. *K.*
61 *pneumoniae* colonization is implicated in chronic diseases of the gastrointestinal tract
62 including inflammatory bowel disease and colorectal cancer[7]. There is also a growing
63 body of evidence highlighting colonization as a reservoir for extraintestinal infections
64 (urinary tract infection, pneumonia, wound or surgical site infections, sepsis) in
65 vulnerable individuals such as neonates, the elderly, immunocompromized and
66 hospitalized patients[8]. Treatment of healthcare-associated (HA) *K. pneumoniae*
67 infections is often limited by multidrug resistance (MDR) resulting from the
68 accumulation of horizontally-acquired AMR genes and mutations in core genes[2].
69 Treatment is further complicated by increasing frequencies of strains producing
70 extended-spectrum β-lactamases (ESBL) and/or carbapenemases, prompting
71 increased reliance on colistin and β-lactamase inhibitor combinations[9,10]. The World
72 Health Organization has accordingly prioritized *K. pneumoniae* as a target for new
73 drugs and therapies[11].

74

75 Outside healthcare settings, *K. pneumoniae* is also recognized as a causative agent of
76 community-acquired infections including urinary tract infection and pneumonia, but
77 also invasive infections such as pyogenic liver abscess, endophthalmitis and
78 meningitis[12,13]. Invasive community-acquired infections are generally associated with

79 so-called hypervirulent *K. pneumoniae* (hvKp) and are most commonly reported in
80 East and Southeast Asia, or in individuals with East Asian ancestry[12]. Features
81 associated with hvKp include a K1, K2 or K5 polysaccharide capsule and
82 horizontally-acquired virulence factors encoding the siderophores aerobactin (Iuc)
83 and salmochelin (Iro), the genotoxin colibactin (Clb), and a hypermucoid phenotype
84 (conferred by the *rmpADC* locus)[14–18]. HvKp are rarely MDR and most remain
85 susceptible to drugs except ampicillin, to which *K. pneumoniae* are intrinsically
86 resistant due to the chromosomally-encoded β-lactamase SHV[19]. However there have
87 been increasing reports of hvKp carrying AMR plasmids and co-occurrence of AMR
88 and virulence determinants in non-hvKp isolates. The convergence of AMR and
89 virulence in *K. pneumoniae* potentiates invasive and difficult-to-treat infections, and
90 at least one fatal outbreak has been documented in China where carbapenemase-
91 producing hvKp are increasingly common[20–24].
92
93 Research conducted in the pre-genomic era characterized 77 distinct capsular (K)
94 serotypes[25], nine O types[26] and variable AMR profiles amongst the *K. pneumoniae*
95 population[27,28], indicating a diverse genetic and phenotypic landscape[15,29]. In recent
96 years, genomic studies have provided key insights into the population structure of *K.*
97 *pneumoniae* (recently summarized in Wyres et al[16]), revealing hundreds of deep-
98 branching phylogenetic lineages comprising sequence types (STs) or clonal groups
99 (CGs) defined by the seven-gene multi-locus sequence typing (MLST) scheme[29].
100 Some of these lineages correspond to lineages (i.e. STs and CGs) that have
101 accumulated large numbers of AMR genes that have become globally distributed (e.g.
102 CG258, CG15, ST307); these are dubbed MDR clones and have been linked with HA
103 infections and hospital outbreaks worldwide[30]. Others carry a high load of virulence
104 genes (e.g. CG23, CG65, CG86) and are recognized as hvKp associated with

105 community-acquired infections. Further distinguishing MDR from hvKp clones are
106 their K and O antigen profiles, with the former displaying a diverse range of K and O
107 biosynthesis loci as a result of homologous recombination between strains, while
108 hvKp rarely deviate from the K1, K2 or K5 types[16].
109
110 Importantly, genomic characterization of clinical isolates identified as *K. pneumoniae*
111 via biochemical tests or mass spectrometry (MALDI-TOF) has revealed the existence
112 of multiple related species and subspecies, which together form the *K. pneumoniae*
113 species complex (KpSC). These differ by 3-4% nucleotide divergence across core
114 chromosomal genes, but share the same pool of AMR and virulence genes[16].
115 Infections and outbreaks caused by other KpSC members have been reported but they
116 generally account for a significantly lower disease burden than *K. pneumoniae* (10-
117 20%)[19,31,32]. Genomics has also clarified that the two *K. pneumoniae* subspecies
118 originally defined by distinct and unusual disease manifestations (subsp.
119 *rhinoscleromatis* which causes a progressive and chronic granulomatous infection
120 known as rhinoscleroma, and subsp. *ozaenae* which causes atrophic rhinitis or ozena)
121 actually represent CGs of *K. pneumoniae* (CG3 and CG90)[15]. Like hvKp clones, these
122 strains also express specific capsule types (K3, K4 and K5) alongside aerobactin and
123 another acquired siderophore, yersiniabactin (Ybt)[16].
124
125 Due to its clinical importance and increasing AMR, *K. pneumoniae* is increasingly the
126 focus of surveillance efforts and molecular epidemiology studies. The sheer volume
127 of clinically-relevant molecular targets renders whole genome sequencing (WGS) the
128 most cost-efficient characterization approach, however extracting and interpreting
129 clinically important features is challenging. To address this, we have developed
130 Kleborate, a genotyping tool designed specifically for *K. pneumoniae* and the

131 associated species complex, which consolidates detection and genotyping of key
132 virulence and AMR loci alongside species, lineage (ST) and predicted K and O
133 antigen serotypes directly from genome assemblies. Here we describe Kleborate and
134 demonstrate its utility by application to publicly available datasets. First, we show
135 that Kleborate can rapidly recapitulate and augment the key findings from a recent
136 large-scale European genomic surveillance study[33]. Next, we apply Kleborate to a
137 curated collection of 13,156 publicly available WGS to further showcase its utility
138 and derive novel insights into the global epidemiology of *Klebsiella* AMR, virulence
139 and convergence. Finally, we show that Kleborate can also be applied to detect
140 clinically relevant genotypes from meta-genome assembled genomes (MAGs).
141
142
143 **RESULTS**
144
145 **Integrated genomic framework and genotyping tool**
146 Our goal was to develop a single tool that can rapidly extract genotype information
147 that is clinically relevant to *K. pneumoniae* and other members of the species complex
148 in order to support genomic epidemiology and surveillance. We have previously
149 reported genotyping schemes for the acquired *K. pneumoniae* virulence loci *ybt, clb,*
150 *iuc* and *iro*[34,35] (whose detection and typing is implemented in early versions of
151 Kleborate), and also K and O antigen typing implemented in the software Kaptive[36].
152 Here we expand the Kleborate framework to include additional features including
153 taxonomic assignment to species and subspecies, assignment to lineages via seven-
154 locus MLST, detection and genotyping of the *rmp* hypermucoidy locus and the *rmpA2*
155 gene, and identification of AMR determinants (mutations and horizontally acquired
156 genes, including assignment of SHV β-lactamase alleles as either ESBL, β-lactamase

157 inhibitor resistance, or intrinsic ampicillin resistance only, see **Methods** and
158 **Supplementary text)**. Kleborate can optionally call Kaptive for K/O antigen
159 prediction.
160
161 Unlike generic AMR or virulence typing tools, we include only genetic features for
162 which there is strong evidence of an associated phenotype in *K. pneumoniae* that has
163 confirmed clinical relevance. These are reported in a manner that facilitates
164 interpretation, including summarizing virulence and AMR genotypes into scores that
165 reflect escalating clinical risk in *K. pneumoniae* infections. Kleborate features are
166 summarized in **Table 1** and methodological details for genotyping are provided in
167 **Methods**. For a typical 5.5 Mbp genome, a Kleborate run including AMR typing
168 takes <10 seconds on a laptop, while robust K and O serotype prediction using
169 Kaptive[36] adds an additional ~1 minute. Results are output in tab-delimited format,
170 making it easy to integrate Kleborate into existing workflows.
171
172 ***Species and subspecies assignment***
173 The taxonomy of *Klebsiella* is rapidly evolving, with several new species and
174 subspecies recently identified[37–39]. As a consequence, many genomes in public
175 databases are incorrectly assigned. We therefore introduced a custom approach for
176 rapid and accurate species and subspecies identification for *Klebsiella*, based on Mash
177 distances[40] to a taxonomically-curated genome set (representative tree in **Figure S1A-**
178 **B**), avoiding the need for users to download large reference genome databases (see
179 **Methods**). This approach was validated using a set of n=285 diverse clinical isolates
180 and compared with species assignments based on the read-based taxonomic classifier
181 Kraken2 (details in **Supplementary Text**, **Table S1, Figure S1**).
182

183 ***Virulence and AMR scores***

184 Genomes are scored according to the clinical risk associated with the AMR and
185 virulence loci that are detected (see **Methods**). Here we take advantage of the
186 structured distribution of AMR and virulence determinants within the *K. pneumoniae*
187 population[14] to reduce the genotyping data to simple numerical summary scores that
188 reflect the accumulation of loci contributing to clinically relevant AMR or
189 hypervirulence: virulence scores range from 0 to 5, depending on the presence of key
190 loci associated with increasing risk (yersiniabactin < colibactin < aerobactin);
191 resistance scores range from 0 to 3, based on detection of genotypes warranting
192 escalation of antimicrobial therapy (ESBL < carbapenemase < carbapenemase plus
193 colistin resistance, see **Table 1**). These simple numerical scores facilitate downstream
194 analyses including trend detection. For example, analysis of a non-redundant subset of
195 9,705 publicly available *K. pneumoniae* genomes (see below, **Table S2**) showed
196 increasing AMR and virulence scores over time (barplots in **Figure 1A-B**). The
197 virulence and resistance scores were correlated not only with the prevalence of
198 individual components that contribute to the scores, but also with other components
199 that are co-distributed in the population (lines in **Figure 1A-B**). For example, the
200 frequencies of *rmpADC* and *rmpA2* loci over time were correlated with the virulence
201 score (**Figure 1A**); and the resistance score was correlated with the mean number of
202 acquired AMR genes and associated drug classes (excluding ESBLs, carbapenemases
203 and colistin which contribute to the score) (**Figure 1C**). Consistent with this, genomes
204 with resistance scores >0 (assigned based on the presence of ESBL and/or
205 carbapenemase genes) typically carry many additional AMR genes conferring
206 resistance to multiple drug classes (**Figure 1D-E**). Reducing the data to key axes of
207 virulence and AMR also facilitates exploration of subpopulations associated with

208 AMR, virulence or convergence of both traits; such as specific *K. pneumoniae*
209 lineages or specimen types (see below).
210
211 **Rapid genotyping of clinical isolates from a large-scale surveillance study**
212 We applied Kleborate to analyse all *K. pneumoniae* clinical isolate genomes deposited
213 in RefSeq by the EuSCAPE surveillance study (927 carbapenem-non-susceptible, 697
214 carbapenem-susceptible; see **Table S2**)[33]. Kleborate rapidly and accurately
215 reproduced key findings from the original study, which were originally derived from
216 multi-step analyses comprising five independent tools and four independent databases
217 (each from a different public repository, one with additional manual curation): (i)
218 70.2% of carbapenem-non-susceptible genomes (n=651/927) carried carbapenemases,
219 mainly KPC-3, OXA-48, KPC-2 and NDM-1; (ii) these were dominated by a few
220 major clones, ST11, ST15, ST45, ST101, ST258, and ST512; (iii) individual countries
221 were associated with specific carbapenemase/clone combinations (see **Figure 2A**). A
222 detailed comparison of the results reported by Kleborate versus those reported in the
223 original study is provided in **Supplementary Text** and **Table S3**.
224
225 In addition to the detection of carbapenemase genes, Kleborate also identified porin
226 defects, which are known to contribute to the carbapenem-resistance phenotype[41,42],
227 in 36.5% of EuSCAPE genomes (including 60% of those with carbapenemase genes
228 and 19.9% of those without carbapenemase genes). These defects included
229 truncation/deletion of OmpK35 and/or OmpK36 (also considered in the original
230 study) as well as GD or TD insertions in the OmpK36 β-strand loop[41] (not considered
231 in original study, but here detected in 18.6% of genomes including 18 with no porin
232 deletion). **Figure 3** shows meropenem MICs stratified by porin defect-carbapenemase
233 combinations identified by Kleborate, highlighting the importance of porin defects –

234 including the OmpK36 β-strand loop insertions – for full expression of carbapenem
235 resistance in *K. pneumoniae*.
236
237 The rise in carbapenem-resistant *K. pneumoniae* infections in hospitals and its
238 associated morbidity and mortality[43] has led to increased interest in alternative control
239 strategies such as vaccines, phage therapy and antibody therapy, key targets for which
240 are the K and O surface antigens[44,45]. Kleborate confidently identified K and O
241 biosynthesis loci in 98.3% and 99.1% of EuSCAPE genomes, respectively, including
242 87 distinct K loci and 11 distinct O loci (**Figures S2** and **S3**). Amongst carbapenem-
243 non-susceptible isolates (meropenem MIC >2), 38 distinct K types were identified
244 and the most common were KL107 (n=173), KL17 (n=67), KL106 (n=41), KL24
245 (n=35), KL15 (n=19) and KL36 (n=13). Seven distinct O types were detected among
246 these genomes, and the most common were O2 (n=294), O1 (n=136) and O4 (n=52).
247 Overall, the data suggest an intervention would need to be effective against six K
248 types or two O types in order to provide coverage of 80% of carbapenem-resistant
249 infections in Europe (**Figure 2B-C**). However, it is important to explore the impact of
250 population structure on these findings, specifically the impact of local clonal
251 expansions. Kleborate aides this type of analysis by providing ST and other
252 genotyping information alongside the K and O locus types, which can be viewed in
253 the context of geographic information. Doing so revealed that each of the top three K
254 loci were dominated by a single ST (83.5% of KL107 were ST512; 93.0% of KL105
255 were ST11; 91.4% of KL17 were ST101). Importantly, the vast majority of ST512-
256 KL107 genomes (75.3%) originated from Italy where this ST is known to be locally
257 circulating[46,47], while 58% of ST11-KL105 originated from Poland and Slovakia, and
258 56% of ST101-KL17 originated from Serbia and Romania. When these putative local

259 expansions were excluded, the top 6 K loci were (KL24, KL15, KL2, KL112, KL107,
260 KL151) and accounted for just 34% of the remaining genomes.
261
262 **Global population snapshot of *K. pneumoniae* AMR and virulence**
263 We applied Kleborate to analyse n=13,156 *Klebsiella* genomes (see **Methods**, **Table**
264 **S2**). Here we provide a brief overview of the data followed by an exploration of
265 AMR, virulence and the phenomenon of convergence, with the aim to highlight the
266 rich information and types of inferences that can be derived from Kleborate output.
267
268 The genome data represented isolates collected from a range of sources in 99
269 countries between 1920–2020 (**Table S4**, although human isolates from the USA,
270 China and UK dominated the data set accounting for n=4,702 genomes, 35.7% of
271 total). The majority of these genomes were sourced from RefSeq, and among these
272 Kleborate identified 1.0% (n=103/10,747) as a species other than the taxon recorded
273 in NCBI; this is consistent with other studies and highlights the current confusion
274 around taxonomic designations in *Klebsiella*. The most common species was *K.*
275 *pneumoniae* (n=11,259, 86%); the rest comprised other KpSC species (9.4%), other
276 members of the *K. oxytoca* species complex (3.1%) and *K. aerogenes* (1.9%) (**Figure**
277 **4, Table S4**). AMR and virulence genes were concentrated in the KpSC and
278 particularly *K. pneumoniae* (**Figure 4, Table S5**).
279
280 The collection captured extensive phylogenetic diversity across the *K. pneumoniae*
281 species (see interactive phylogeny at
282 http://microreact.org/project/JDyan46yctyDh6weEUjWN), and Kleborate assigned
283 these genomes to ≥1,452 different STs (1,119 known STs across and at least 333
284 novel STs). Notably, 600 STs (41%) were represented by just a single genome each

285 (accounting for 5.3% of all genomes). We detected n=4 ST67 (subspecies
286 *rhinoscleromatis*) and n=3 ST90 (subspecies *ozanae*). A small number of STs were
287 overrepresented, reflecting the bias towards sequencing MDR and hypervirulent
288 isolates, as well as those causing hospital outbreaks. For example, 1,354 genomes
289 (12.0%) represented the KPC-associated ST258, which is known to dominate
290 carbapenem-resistant *K. pneumoniae* in the USA and southern Europe (where it has
291 been the subject of intense genomic investigations) but is comparatively rare in other
292 regions of the world[16]. To reduce the impact of these sampling biases in public
293 genome collections, we down-sampled to a non-redundant set of 9,705 *K.*
294 *pneumoniae* genomes representing unique combinations of ST, genetic subcluster
295 (Mash distance <0.0003), virulence genotype, AMR genotype, specimen type,
296 location and year of isolation (see **Methods**). However, we cannot fully correct for
297 the sampling biases inherent in the public genome data and even after subsampling,
298 the 30 most common STs accounted for 63.4% of genomes (n≥50 genomes each,
299 n=6,151 total; see **Figure S4**). **Figure 5** shows the distribution of AMR and virulence
300 scores amongst non-redundant genomes from these 30 common *K. pneumoniae* STs
301 (n>50 per ST), each of which displays high rates of AMR and/or virulence.
302
303 ***AMR determinants***
304 SHV β-lactamases conferring intrinsic resistance to the penicillins were detected in
305 85.9% of the 9,705 non-redundant *K. pneumoniae* genomes (ESBL forms of SHV
306 were detected in 10.0%). Acquired AMR was widespread (77.1% of genomes had at
307 least one gene or mutation conferring acquired AMR detected) and 71.6% of genomes
308 were predicted to be MDR (acquired resistance to ≥3 drug classes[48]), a much higher
309 rate than is reported in most geographical regions[3,49–51], reflecting the bias within
310 public genome collections. The majority of genomes had a non-zero resistance score,

311 reflecting presence of ESBL and/or carbapenemase genes: 22.3%, 37.1% and 5.9%
312 genomes had resistance scores of 1, 2 and 3 respectively. Mean resistance scores
313 increased through time (**Figure 1B**). This trend could be an artefact of sampling bias
314 towards the selective sequencing of AMR isolates, however it is consistent with the
315 increasing AMR rates reported in surveillance studies globally[52–54].
316
317 Comparatively higher prevalence of acquired AMR genes was observed in some STs
318 (**Figure S4**). Many of these STs represent recognized MDR clones largely from
319 clinical samples that were also associated with high mean resistance scores (**Figures**
320 **6A-B**), driven by high frequency of ESBL and carbapenemase genes (**Figures 5,**
321 **S5A-B**). The most common ESBLs/carbapenemases were widely detected across the
322 population (46-299 STs each), including amongst the top 30 common STs (prevalence
323 range per ST, 0.1-100%; see **Figure S5A-B**), highlighting their mobile nature. The
324 notable exception was CTX-M-65, which appeared to be largely clone specific,
325 detected in only 9 STs and ST11 accounting for 96.7% of these genomes.
326
327 Colistin resistance determinants were detected in 8.7% of the non-redundant *K.*
328 *pneumoniae* genomes. These were mostly nonsense mutations in MgrB or PmrB
329 (83.5%) rather than acquisition of an *mcr* gene (15.8%, and an additional 6 genomes
330 with both acquired *mcr* and truncated MgrB/PmrB). The rate of detection ranged from
331 0-25.2% for the 30 most common STs, and was highest amongst ST512, ST437,
332 ST147, ST16 and ST258 (**Figure S5C)**, each of which are also associated with high
333 rates of carbapenem-resistance. Porin mutations were detected in 37.9% of genomes
334 (34.0% OmpK35, 20.2% OmpK36, 16.3% both). High prevalence of specific porin
335 defects have been reported previously in some clones[41,42], and this was reflected in
336 our analysis of ST258 and its derivative ST512. We observed OmpK35 truncations in

337 99.9% of non-redundant ST258 genomes (with or without truncations or substitutions
338 in OmpK36), and truncations in OmpK35 and/or OmpK36 in all ST512 (99.4% with
339 OmpK35 truncations, 94.4% with the OmpK36GD mutation, see **Figure S5D**).
340
341 ***Virulence loci***
342 The prevalence of acquired siderophores and colibactin loci amongst non-redundant
343 *K. pneumoniae* genomes was 44.4% *ybt*, 7.5% *clb*, 11.2% *iuc* and 7.0% *iro*. The loci
344 were found across diverse *K. pneumoniae* STs (391 STs with *ybt*, 56 with *clb*, 144
345 with *iuc*, 108 with *iro*) but were rarely detected in other *Klebsiella* species (with the
346 exception of *ybt* among the *K. oxytoca* species complex, see **Figure 4**) indicating
347 frequent mobilisation within *K. pneumoniae* but not between species (**Table S6**,
348 **Figure S6**). Mean virulence scores increased through time (**Figure 1A**). **Figure 5B**
349 shows the frequency of virulence scores in the top 30 most common STs in the non-
350 redundant genome set. Sixteen of these common STs had ≥40% of genomes carrying
351 the ICE*Kp*-associated *ybt* without the virulence plasmid-associated *iuc* locus (i.e.
352 virulence score=1-2), including well known MDR clones ST258, ST11, ST14, ST15,
353 ST101, ST147, ST152, ST395. Only the hvKp clones (ST23, ST86, ST65) and ST231
354 had a high frequency of *iuc* (virulence score ≥3).
355
356 In addition to detecting the presence of virulence loci, Kleborate reports on their
357 completeness, genetic lineages and associated MGE variants, which can provide
358 insights into their dissemination. Most of the virulence loci identified in the non-
359 redundant *K. pneumoniae* data set (98%) matched one of the genetic lineages
360 described previously[34,35] (**Table S6**). **Figure S7A** shows the frequency of *iuc* lineages
361 in *K. pneumoniae* STs with ≥20 non-redundant genomes and at least one genome
362 harbouring *iuc*. There were four STs for which >60% genomes harboured *iuc*, and

363 only a single *iuc* lineage was detected in each (*iuc1* in ST23, ST65, ST86; *iuc2A* in
364 ST82), consistent with long-term persistence of a specific virulence plasmid in these
365 well-known hypervirulent clones. In contrast, *iuc* was less frequent among other STs,
366 several of which were associated with multiple *iuc* lineages (e.g. ST231, ST25,
367 ST35), consistent with more recent and/or transient virulence plasmid acquisitions
368 (mostly *iuc1*, followed by *iuc3* and *iuc5*).
369
370 Frameshift mutations (i.e. truncations) and/or incomplete loci (i.e. missing at least one
371 gene) were detected in 10%, 28.5%, 13.6% and 17.7% of non-redundant *K.*
372 *pneumoniae* genomes with *ybt*, *clb*, *iuc* and *iro* respectively (**Table S7**). While some
373 of these may erroneously arise from contig breaks in draft genome assemblies, true
374 truncations or missing genes may reflect a lack of function. The latter is likely true for
375 instances where we observe conserved frameshift mutations across entire lineages,
376 e.g. frameshift mutations were detected in *iucA* for all *iuc3*+ genomes and in *iroC* for
377 all *iro3*+ and *iro4*+ genomes.
378
379 The hypermucoidy locus *rmpADC* was detected in 8.4% of non-redundant *K.*
380 *pneumoniae* genomes (and just eight genomes of other KpSC species, **Table S6**). The
381 majority of these genomes (67.2%, belonging to >79 STs) carried intact copies of all
382 three genes, thus likely express the hypermucoid phenotype. Intact *rmpADC* was
383 common in *iuc*-positive genomes of the hvKp clones ST23 and ST86, as well as MDR
384 clones ST29 and ST101 (**Figure S7B**). Many other *iuc*-positive genomes carried
385 *rmpADC* loci with truncated or missing genes, which likely do not confer the
386 hypermucoid phenotype. Notably, these included hvKp clones ST65 and ST82, as
387 well as MDR clones ST231, ST15 and ST14. The *rmpA2* gene was detected in 7.4%
388 of non-redundant *K. pneumoniae* genomes, but was mostly present in truncated form

389 (89.0% of *rmpA2*+ genomes) due to frameshifts within a poly-G tract[55]. The latter
390 highlights the importance of considering not only the presence/absence of a given
391 gene, but also whether it encodes a full-length protein, which may have important
392 clinical implications.
393
394 **Facilitating detection of AMR-virulence convergence**
395 AMR and virulence determinants have until recently been segregated in non-
396 overlapping *K. pneumoniae* populations[14,19], as clearly indicated by the distributions
397 of AMR and virulence scores among STs (**Figures 5, 6A**). However, reports of
398 convergent AMR-virulent strains with the potential to cause difficult-to-treat
399 infections are increasingly common[16,56]. Kleborate facilitates rapid identification of
400 such strains on the basis of resistance and virulence scores (convergence defined as
401 virulence score ≥3 and resistance score ≥1, **Figure 6C**). Based on these scores, we
402 observed a total of 601 convergent *K. pneumoniae* (510 non-redundant) with the
403 highest proportion corresponding to a virulence score of 4 (indicative of
404 yersiniabactin plus aerobactin/virulence plasmid detection) and resistance score of 2
405 (carbapenem resistance).
406
407 The majority of convergent genomes (74.5%) were concentrated within a small
408 number of STs comprising the well-known hypervirulent (ST23, ST86, ST65) and
409 MDR lineages (ST11, ST15, ST231 and ST147) (**Figures 6C-D, 7**). We combined the
410 genotyping data and information from a Mash-distance-based neighbour-joining tree
411 (http://microreact.org/project/JDyan46yctyDh6weEUjWN) to define unique
412 convergence events (defined as unique combinations of ST, virulence and resistance
413 determinants, and phylogenetic cluster). This identified n=173 convergence events,
414 accounted for by either acquisition of the virulence plasmid by MDR/other clones

415 (n=84 events; 475 genomes), or acquisition of ESBLs/carbapenemases by
416 hypervirulent clones (n=89 events; 126 genomes) (**Figure 7**, **Table S8**).
417
418 The most common virulence plasmid, KpVP-1 (*iuc1* ± *iro1*), accounted for 54% of
419 virulence plasmid acquisition events (n=45 acquisitions), while *iuc3* plasmids, the *E.*
420 *coli* derived *iuc5* (±*iro5*) and *iuc/iro* unknown (i.e. novel or divergent *iuc/iro* loci)
421 accounted for 21%, 11% and 14%, respectively (**Figure 7**). AMR acquisitions by
422 hypervirulent clones involved the ESBL/carbapenemase genes that are most common
423 in the general *K. pneumoniae* population: KPC-2 (26%), OXA-232 (17%) and CTX-
424 M-15 (18%). The majority of convergence events (87%) were associated with just a
425 small number of genomes (i.e. n≤3); however, five events were associated with >20
426 genomes in the complete dataset, which may indicate clonal expansion and
427 dissemination of the corresponding convergent strains locally and/or between
428 countries. One such event corresponded to the ST11-KPC + KpVP-1 deletion variant
429 strain that was originally reported in 2017[20] and has since been recognized as widely
430 distributed in China[20–24]. The complete public genome set (i.e. counting redundant
431 genomes) included 148 genomes corresponding to this specific ST11 convergence
432 event mostly from China but also from France (n=2). Notably though, this was only
433 one of 50 convergence events that we detected in China, including 8 involving
434 acquisition of *iuc1* or *iuc5* by ST11 (see **Table S8**, and interactive tree at
435 http://microreact.org/project/JDyan46yctyDh6weEUjWN). Additional events associated
436 with >20 genomes included (i) ST231-MDR + virulence plasmids carrying novel *iuc*
437 lineages detected in India, Pakistan, Switzerland, Thailand and USA, (ii) ST15-CTX-
438 M-15 + KpVP-1 in Pakistan, (iii) ST15-MDR + KpVP-1 in China and Nepal, and (iv)
439 another distinct ST11-KPC-2 + KpVP-1 event in China. Including the above three

440 examples, 11 convergence events appeared to involve intercountry expansion of
441 which one has been previously documented[57].
442
443 Overall, convergent genomes were detected originating from most geographical
444 regions for which genome data was available, but some regions had many more
445 events than others (**Figure 7, Table S8**). This uneven distribution may stem from a
446 skew in the number of genomes available per region (e.g. due to variation in
447 accessibility or application of genome sequencing). Nevertheless, the number of
448 convergent genomes in the eastern, southeastern and southern parts of Asia were
449 noticeably high, driven by the frequency of convergence events detected in China
450 (n=50 events) and Thailand (n=26 events) as well as putative clonal expansions of
451 these strains as discussed above (**Figure 7**). Of note, AMR acquisitions by
452 hypervirulent lineages were particularly frequent within East and Southeast Asia
453 where hypervirulent infections are most frequently reported, alongside countries from
454 eastern and northern Europe.
455
456 Outside of *K. pneumoniae,* convergence events were rare: we detected n=2 *K.*
457 *quasipneumoniae* subsp. *similipneumoniae* (ST367 with KpVP-1 and CTX-M-15;
458 ST3387 with *iuc3* and CTX-M-55) and n=2 *K. variicola* subsp. *variicola* (ST595 with
459 KpVP-1 and KPC-2; ST1848 with *iuc5* and KPC-2).
460
461 **Genotyping *K. pneumoniae* from metagenome data**
462 There is increasing interest in detection and typing of *K. pneumoniae* direct from gut
463 metagenome data[58], due to the role of *K. pneumoniae* gut colonization as a source of
464 acute infections and as a contributor to chronic diseases[7,8]. We tested Kleborate's
465 performance by application to n=40 metagenomes from which at least one KpSC

466 isolate was cultured and sequenced, as part of the Baby Biome Study[59]. We compared
467 the results of running Kleborate on metagenome-assembled genomes (MAGs, i.e.
468 species-specific contig bins extracted from whole-metagenome assemblies) vs. KpSC
469 isolate whole genome sequence(s) cultured from the same fecal sample. Thirty-two
470 metagenomes had >1% relative abundance of KpSC, and genotyping of MAGs from
471 these yielded results consistent with genotyping of cultured isolates for 26/32 samples
472 (16 with identical genotypes reported for species, ST, K/O locus, virulence and AMR;
473 10 with close matches; see **Fig. S8, Tables S9-S10**). As expected, MAG-derived
474 genotypes were closest to those of isolates when only one KpSC strain was cultured
475 from the sample (see **Fig. S8, Table S10**). Kleborate analysis of whole metagenome
476 assemblies (as opposed to individual MAGs) is not recommended: species detection
477 and ST assignment matched that of the corresponding WGS isolates for only n=4/40
478 metagenome assemblies, which is unsurprising as the whole metagenomes include
479 sequences derived from dozens of different bacteria, many of which harbour
480 homologs of genotyping targets.
481
482 **DISCUSSION**
483 Whole genome sequencing is being increasingly implemented in research and public
484 health labs as a cost- and time- effective option for tracking pathogens and AMR.
485 However, identification of clinically-relevant features remains a key bottleneck that
486 hinders widespread adoption of genome surveillance. We have presented a
487 comprehensive framework and tool for rapid genotyping of *Klebsiella* species
488 genomes: Kleborate is a single unified approach for species detection, MLST and
489 genotyping of key virulence and AMR determinants. It focusses only on genomic
490 features for which there is strong evidence of a clinically relevant phenotype in KpSC

491 and presents the data in a readily interpretable format, with numerical summaries and
492 categorical scores corresponding to measures of clinical risk.
493
494 A key strength of the Kleborate framework is its species-specific approach. This is
495 particularly important for accurate interpretation of AMR and virulence gene screens
496 from WGS, wherein the use of generic databases and tools can result in confusion.
497 Notable examples include the intrinsic *oqxAB* and *fosA* alleles, which unlike for other
498 Enterobacterales, do not confer resistance to quinolones and fosfomycin when
499 expressed in KpSC. Kleborate does not report these intrinsic alleles, neither does it
500 report intrinsic virulence determinants such as the siderophore enterobactin, which is
501 known to play a role in KpSC pathogenicity but for which the presence alone cannot
502 be considered to indicate enhanced virulence of one isolate over another. Correct
503 taxonomic identification of *K. pneumoniae* can be difficult in itself, hence the inbuilt
504 speciation tool is an important feature (and here identified nearly 100 RefSeq
505 genomes with incorrect species/subspecies assignments).
506
507 Another strength of our approach is the rich data output by Kleborate, which
508 facilitates in-depth investigation of population structure, AMR and virulence
509 epidemiology. This allows rapid exploration and understanding of: (i) hypervirulence-
510 associated loci and the molecular drivers of their dissemination (**Figures S4** and **S7**);
511 (ii) molecular mechanisms of complex AMR phenotypes e.g. carbapenem resistance
512 (**Figure 3**); (iii) AMR and virulence trends (**Figures 1**, **5** and **6**); (iv) emerging
513 convergent AMR-virulent strains so that they can be targeted for surveillance and
514 infection control (**Figure 7**); (v) overrepresented STs and genotypes, which may be
515 indicative of transmission clusters that should be targeted for further investigation (as
516 demonstrated for the EuSCAPE surveillance genomes, **Figure 2A**); (vi) surface

517 antigen epidemiology, which can inform the design of novel vaccines and
518 therapeutics (**Figure 2B-C**). Notably Kleborate can also yield useful genotyping
519 results from metagenomics data (**Figure S8**), which is gradually being adopted for
520 clinical and surveillance applications relevant to *K. pneumoniae*. User interpretation
521 of Kleborate's extensive data output can be guided by the accompanying web-based
522 visualization app, Kleborate-Viz. Through this app, many of the analyses and plots
523 presented in this manuscript can be rapidly replicated, and further explored in an
524 interactive manner.
525
526 Kleborate is designed to facilitate detection and tracking of clinically relevant AMR
527 and virulence traits from genome data, and analysis of public data not only identified
528 specific clones and genes associated with one or the other of these traits (**Figures 5,**
529 **6**), but also 601 genomes in which the two converge (carrying *iuc*+ virulence
530 plasmids and ESBL and/or carbapenemase genes; **Figure 7**). We estimated at least
531 173 unique AMR-hypervirulence convergence events; the majority were detected
532 within a single isolate (n=119 events), but many others appear to be associated with
533 local outbreaks or larger-scale spread and apparently across multiple countries (**Table**
534 **S8**). Some of the convergence events in China and other countries in the neighbouring
535 South and Southeast Asia regions have been extensively reported[16,20,49,56], but to our
536 knowledge a significant number had not been recognized previously. These include
537 ST231-MDR (most with OXA-232, remainder with ESBL only) + *iuc*, which has
538 been reportedly circulating in India[49], and our analysis also detected in Pakistan,
539 Thailand, Switzerland and USA.
540
541 Kleborate has already been widely adopted by the *Klebsiella* research community – at
542 least 74 studies have reported using the Kleborate software package, including larger-

543 scale genome surveillance studies in South and Southeast Asia, the Caribbean and the
544 United States[31,49,50] (full list in **Table S11**). Kleborate is freely available as a
545 standalone command-line tool for local high-throughput analyses or incorporation
546 into existing bioinformatics workflows (https://github.com/katholt/Kleborate), and
547 can be easily accessed through the online tool PathogenWatch
548 (https://pathogen.watch/). With such broad accessibility and utility, Kleborate is
549 poised to become a cornerstone of the *Klebsiella* genomic surveillance toolkit that can
550 help inform containment and control strategies targeting this priority pathogen.
551
552
553 **METHODS**
554
555 **Kleborate software: implementation and genotyping logic**
556 Kleborate (v.2) is a command-line tool written in Python and is freely available under
557 the GNU v3.0 license at http://github.com/katholt/Kleborate. It takes as input one or
558 more whole genome assemblies (FASTA format), types each one against a series of
559 screening databases outlined in detail below, and returns results in a tab-delimited text
560 file (one genome per row). On default settings, Kleborate will report assembly quality
561 metrics, taxonomic assignment, MLST and virulence loci genotypes. Screening for
562 AMR determinants, and/or K/O serotyping via Kaptive[36], is optional (**Table 1**).
563
564 ***Assembly quality***
565 Assembly quality metrics, reported to help users assess the reliability of genotyping
566 results, are: contig count, contig N50, largest contig size, total genome size, and
567 number of ambiguous bases (e.g. 'N'). Low quality warnings are flagged if: (i)
568 ambiguous bases are detected; (ii) assembly length falls outside the expected range of

569 4.5-7.5 Mbp; or (iii) N50 is below 10,000 bp. Users should carefully consider the
570 genotyping outputs for low quality assemblies.
571
572 ***Taxonomic assignment***
573 Kleborate's species prediction function provides a convenient way to confirm species,
574 including differentiating between the closely related members of the KpSC which are
575 frequently misclassified using laboratory techniques. Kleborate calculates Mash[40]
576 distances between the input genome/s and a curated collection of reference assemblies
577 from different *Klebsiella* and other Enterobacterales, and reports the species with the
578 smallest distance. Mash distance ≤0.02 is reported as a strong match, ≤0.04 as weak
579 (only when no strong matches are found, see **Supplementary Text** for further
580 details).
581
582 ***MLST***
583 Genomes assigned to species in the KpSC are assigned sequence types using
584 nucleotide BLAST against the established *K. pneumoniae* chromosomal seven-locus
585 MLST scheme[29] described and maintained on the *K. pneumoniae* BIGSdb site hosted
586 at the Pasteur Institute (http://bigsdb.pasteur.fr/klebsiella/klebsiella.html).
587
588 ***Virulence gene detection and typing***
589 Virulence loci (*ybt*, *iuc*, *iro*, *clb*, *rmpADC, rmpA2*) are detected using nucleotide
590 BLAST search against the database of known alleles. The best hit allele for each gene
591 (with ≥90% identity and ≥80% coverage) is reported in the main virulence columns.
592 If the majority of genes expected for the locus are present, then the alleles are used to
593 calculate STs which are reported along with their associated lineage and MGE (based
594 on previously defined schemes: YbST for *ybt*, CbST for *clb*, AbST for *iuc*, SmST for

595 *iro*, according to the previously defined schemes[34,35]; and a novel RmST scheme for
596 the *rmpADC* locus). To generate the RmST typing scheme we used the same 2,733
597 genomes from our original virulence plasmid study[35] to screen and extract the
598 sequences for *rmpADC* and define allele numbers and STs. These ST sequences
599 cluster into four distinct lineages associated with distinct MGEs (*rmp1* with KpVP-1,
600 *rmp2* with KpVP-2, *rmp2A* with the *iuc2A* virulence plasmids, and *rmp3* with
601 ICE*Kp1*; to be described in detail elsewhere). Where the best hit for a gene is a weak
602 match (80-90% identity, 40-80% coverage) this is reported in the 'spurious hits'
603 column. Truncations are detected by translating the best-matching nucleotide
604 sequence for each query gene into amino acids and comparing to the reference length
605 (expressed as % amino acid length from the start codon, those <90% are reported).
606 The presence of *ybt*, *clb* and *iuc* are used to assign a virulence score as follows:
607 0=none present, 1=yersiniabactin only, 2=colibactin without aerobactin (regardless of
608 yersiniabactin, however *ybt* is almost always present when *clb* is), 3=aerobactin only,
609 4=aerobactin and yersiniabactin without colibactin, and 5= all three present. The
610 presence of *iro* (salmochelin) is not used to calculate the virulence score because its
611 presence is very strongly associated with aerobactin.
612
613 ***Detection and typing of antimicrobial resistance determinants***
614 When AMR detection is switched on, Kleborate screens for known acquired AMR
615 determinants using a curated version of the CARD AMR nucleotide database (v3.0.8
616 downloaded February 2020; see doi.org/10.6084/m9.figshare.13256759.v1 for full
617 details on curation). Genes are identified using nucleotide BLAST (and amino acid
618 search with tBLASTx if no exact nucleotide match is found). Gene truncations and
619 spurious hits are identified as described above for virulence genes. Unlike the
620 acquired forms, the intrinsic variants of *oqxAB,* chromosomal *fosA* and *ampH* are not

621 associated with clinical resistance in KpSC and are therefore not reported. However,
622 SHV, LEN or OKP β-lactamase alleles intrinsic to KpSC species are known to confer
623 clinical resistance to penicillins and are reported in the Bla_chr column. Acquired
624 SHV variants, and individual SHV sequence mutations known to confer resistance to
625 extended-spectrum β-lactams or β-lactamase inhibitors, are reported separately (see
626 **Supplementary Text, Tables S12-S13** for details).
627
628 Chromosomally encoded mutations and gene loss or truncations known to be
629 associated with AMR are reported for genomes identified as KpSC species. These
630 include fluoroquinolone resistance mutations in GyrA (codons 83 and 87) and ParC
631 (codons 80 and 84), and colistin resistance from truncation or loss of MgrB and PmrB
632 (defined as <90% amino acid sequence coverage). Mutations in the OmpK35 and
633 OmpK36 osmoporins reportedly associated with reduced susceptibility to β-
634 lactamases[41,42] are also screened and reported for KpSC genomes, and include
635 truncation or loss of these genes and OmpK36GD and OmpK36TD transmembrane β-
636 strand loop insertions[41]. SHV β-lactamase, GyrA, ParC and OmpK mutations are
637 identified by alignment of the translated amino acid sequences against a reference
638 using BioPython, followed by interrogation of the alignment positions of interest (see
639 **Supplementary Text, Tables S12-S13** for a list of relevant positions).
640
641 AMR genes and mutations are reported by drug class, with β-lactamases further
642 categorized by enzyme activity (β-lactamase, ESBL or carbapenemase, with/without
643 resistance to β-lactamase inhibitors). Horizontally acquired AMR genes are reported
644 separately from mutational resistance and contribute to the AMR gene count; these
645 plus chromosomal mutations count towards the number of acquired resistance classes
646 (intrinsic SHV alleles, reported in Bla_chr column, are not included in either count).

647 Resistance scores are calculated as follows: 0=no ESBL or carbapenemase, 1=ESBL
648 without carbapenemase (regardless of colistin resistance); 2=carbapenamase without
649 colistin resistance (regardless of ESBL); 3=carbapenemase with colistin resistance
650 (regardless of ESBL).
651
652 ***Serotype prediction***
653 By default, genomes are screened against the *wzi* database in the *Klebsiella* BIGSdb
654 (using nucleotide BLAST) which is used to predict capsule (K) type based on a
655 previously defined scheme[60]. This allows rapid typing however the relationship
656 between *wzi* allele and K type is not one-to-one[36] . If surface antigen prediction is
657 important to users they can obtain more robust identification of K and O antigen
658 (LPS) loci by switching on serotype prediction with Kaptive[36] (--kaptive), which adds
659 a few minutes per genome to Kleborate's runtime.
660
661 ***Data visualization***
662 To facilitate interpretation of Kleborate's rich data output we provide a web-based
663 application (Kleborate-Viz, https://kleborate.erc.monash.edu/), implemented in R
664 Shiny, which takes as inputs a Kleborate results file (required), sample metadata
665 (CSV format, optional) and MIC data (CSV format, optional).
666
667 **Genome analysis**
668 The analyses reported here result from applying Kleborate v2.0.0 to publicly available
669 genome collections. A total of 13,156 *Klebsiella* WGS assemblies, encompassing
670 non-duplicate isolates with unique BioSample accessions identified from published
671 studies (some deposited as read sets only, which were assembled using Unicycler
672 v0.4.7[61], data sources summarized in **Table S14**) plus any additional genomes

673 designated as *Klebsiella* in NCBI's RefSeq repository of genome assemblies (as of
674 17th July 2020). In order to minimize the impact of sampling bias favouring common
675 MDR and/or virulent lineages and those causing outbreaks, we subsampled the
676 collection into a 'non-redundant' dataset of 11,277 genomes (9,705 *K. pneumoniae*)
677 as follows. Pairwise Mash distances were calculated using Mash v2.1, and used to
678 cluster genomes using single-linkage clustering with a threshold of 0.0003. These
679 clusters were further divided into non-redundant groups with unique combinations of
680 (i) Mash cluster, (ii) chromosomal ST, (iii) virulence gene profiles (i.e. presence of
681 *ybt/clb/iro/iuc* loci and lineage assignment), (iv) AMR profiles, (v) year and country
682 of isolation, and (vii) specimen type where available. For each resulting non-
683 redundant group, one genome was selected at random as the representative for
684 analyses. The full list of genomes, including database accessions, isolate information,
685 cluster/group assignment, and Kleborate results are provided in **Table S2**. The subset
686 of 1,624 *K. pneumoniae* assemblies deposited in RefSeq by the European EuSCAPE
687 surveillance study[33] (out of 1,649 reported in original study; **Table S2**) were used for
688 the EuSCAPE analyses reported in **Figures 2** and **3**. The Kleborate-Viz web
689 application is pre-loaded with the non-redundant and EuSCAPE WGS datasets
690 reported in this paper, and can be used to reproduce the plots shown in **Figures 1A-C,**
691 **2B-C, 3, 6A-B** and to further explore the Kleborate results.
692
693 **Metagenome analysis**
694 We downloaded metagenomic reads, and matched isolate WGS assemblies, for n=47
695 infant gut microbiota samples deposited by the Baby Biome Study[59]. Metagenome
696 reads were assembled using SPAdes version 3.13.1[62] with the --meta flag and the
697 resulting contigs binned using MaxBin v2.2.7[63]. Seven metagenomes failed to
698 assemble due to memory and compute walltime constraints, hence we report results

699 for 40 samples (**Table S10**). Kleborate was run separately on the full metagenome
700 assemblies, all contig bins (from which the *Klebsiella* bin could then be identified),
701 and the matched WGS assemblies. Metagenomic read sets were also analysed using
702 Kracken 2.0.7[64] and Bracken v2.5[65] (with a custom GDTB release 89 database[66]) to
703 estimate the relative abundance of KpSC reads in each metagenome.
704

705 **Statistical analysis**

706 Statistical analyses and data visualisations were conducted using R v1.1.456. Figures
707 were generated with ggplot v3.2.0 and pheatmap v1.0.12. Correlations between
708 virulence and resistance scores, and the prevalence of virulence and resistance
709 determinants over time, were analysed using Spearman's rank-order correlation (i.e.
710 non-parametric test).
711

712 **Acknowledgements**

713 We thank Prof Sylvain Brisse and the curators of the *K. pneumoniae* BIGSdb at
714 Institut Pasteur for hosting and maintaining the MLST schemes
715 (https://bigsdb.pasteur.fr/klebsiella/klebsiella.html); and Prof David Aanensen and
716 team at the Centre for Genomic Pathogen Surveillance for making Kleborate available
717 online within Pathogenwatch (http://pathogen.watch).
718

719 **Funding**

720 This work was supported by the Bill and Melinda Gates Foundation of Seattle (grant
721 OPP1175797 to K.E.H.) and the Viertel Charitable Foundation of Australia (Senior
722 Medical Research Fellowship to K.E.H.). K.L.W is supported by the National Health
723 and Medical Research Council of Australia (Investigator Grant APP1176192).
724

725

726 **References**

727

728 1. World Health Organisation. *Global Priority List of Antibiotic-Resistant*
729 *Bacteria to Guide Research, Discovery, and Development of New Antibiotics*.
730 (2017).
731 2. Wyres, K. L. & Holt, K. E. *Klebsiella pneumoniae* as a key trafficker of drug
732 resistance genes from environmental to clinically important bacteria. *Curr.*
733 *Opin. Microbiol.* **45**, 131–139 (2018).
734 3. Gorrie, C. L. *et al.* Gastrointestinal carriage is a major reservoir of *K.*
735 *pneumoniae* infection in intensive care patients. *Clin Infect Dis* **65**, 208–215
736 (2017).
737 4. Martin, R. M. *et al.* Molecular epidemiology of colonizing and infecting
738 isolates of *Klebsiella pneumoniae*. *mSphere* **1**, (2016).
739 5. Chung, D. R. *et al.* Fecal carriage of serotype K1 *Klebsiella pneumoniae* ST23
740 strains closely related to liver abscess isolates in Koreans living in Korea. *Eur J*
741 *Clin Microbiol Infect Dis* **31**, 481–486 (2012).
742 6. Lin, Y.-T. *et al.* Seroepidemiology of *Klebsiella pneumoniae* colonizing the
743 intestinal tract of healthy chinese and overseas chinese adults in Asian
744 countries. *BMC Microbiol.* **12**, 13 (2012).
745 7. Kaur, C. P., Vadivelu, J. & Chandramathi, S. Impact of *Klebsiella pneumoniae*
746 in lower gastrointestinal tract diseases. *J. Dig. Dis.* **19**, 262–271 (2018).
747 8. Podschun, R. & Ullmann, U. *Klebsiella* spp. as Nosocomial Pathogens:
748 Epidemiology, Taxonomy, Typing Methods, and Pathogenicity Factors. *Clin*
749 *Microbiol Rev* **11**, 589–603 (1998).
750 9. Petrosillo, N., Taglietti, F. & Granata, G. Treatment Options for Colistin

751 Resistant *Klebsiella pneumoniae*: Present and Future. *J. Clin. Med.* **8**, 934
752 (2019).
753 10. Tooke, C. L. *et al.* β-Lactamases and β-Lactamase Inhibitors in the 21st
754 Century. *J. Mol. Biol.* **431**, 3472–3500 (2019).
755 11. Geneva: World Health Organization. *Prioritization of Pathogens To Guide*
756 *Discovery, Research and Development of New Antibiotics for Drug-Resistant*
757 *Bacterial Infections, Including Tuberculosis*. (2017).
758 12. Shon, A. S., Bajwa, R. P. S. & Russo, T. A. Hypervirulent (hypermucoviscous)
759 *Klebsiella pneumoniae*: a new and dangerous breed. *Virulence* **4**, 107–118
760 (2013).
761 13. Siu, L. K., Yeh, K., Lin, J., Fung, C. & Chang, F. *Klebsiella pneumoniae* liver
762 abscess: a new invasive syndrome. *Lancet Infect Dis* **12**, 881–887 (2012).
763 14. Wyres, K. L. *et al.* Distinct evolutionary dynamics of horizontal gene transfer
764 in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet.* **15**,
765 e1008114 (2019).
766 15. Brisse, S. *et al.* Virulent clones of *Klebsiella pneumoniae*: Identification and
767 evolutionary scenario based on genomic and phenotypic characterization. *PLoS*
768 *One* **4**, (2009).
769 16. Wyres, K. L., Lam, M. M. C. & Holt, K. E. Population genomics of *Klebsiella*
770 *pneumoniae*. *Nat. Rev. Microbiol.* **18**, 344–359 (2020).
771 17. Walker, K. A. *et al.* A *Klebsiella pneumoniae* Regulatory Mutant Has Reduced
772 Capsule Expression but Retains Hypermucoviscosity. *MBio* **10**, e00089-19
773 (2019).
774 18. Walker, K. A., Treat, L. P., Sepúlveda, V. E. & Miller, V. L. The Small Protein
775 RmpD Drives Hypermucoviscosity in *Klebsiella pneumoniae*. *MBio* **11**,
776 e01750-20 (2020).

777 19. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence,
778 and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to
779 public health. *Proc Natl Acad Sci USA* **112**, E3574--81 (2015).
780 20. Gu, D. *et al.* A fatal outbreak of ST11 carbapenem-resistant hypervirulent
781 *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological
782 study. *Lancet Infect Dis* **18**, 37–46 (2018).
783 21. Xu, M. *et al.* High prevalence of KPC-2-producing hypervirulent *Klebsiella*
784 *pneumoniae* causing meningitis in Eastern China. *Infect. Drug Resist.* **12**, 641–
785 653 (2019).
786 22. Dong, N. *et al.* Genome analysis of clinical multilocus sequence Type 11
787 *Klebsiella pneumoniae* from China. *Microb. genomics* **4**, e000149 (2018).
788 23. Wong, M. H. Y. *et al.* Emergence of carbapenem-resistant hypervirulent
789 *Klebsiella pneumoniae*. *Lancet Infect Dis* **18**, (2018).
790 24. Yao, H., Qin, S., Chen, S., Shen, J. & Du, X.-D. Emergence of carbapenem-
791 resistant hypervirulent *Klebsiella pneumoniae*. *Lancet Infect Dis* **18**, (2018).
792 25. Ørskov, I. D. A. & Fife-Asbury, M. A. New *Klebsiella* capsular antigen, K82,
793 and the deletion of five of those previously assigned. *Int J Syst Bacteriol.* **27**,
794 386–387 (1977).
795 26. Trautmann, M. *et al.* O-antigen seroepidemiology of *Klebsiella* clinical isolates
796 and implications for immunoprophylaxis of *Klebsiella* infections. *Clin. Diagn.*
797 *Lab. Immunol.* **4**, 550–555 (1997).
798 27. Elhani, D. *et al.* Molecular epidemiology of extended-spectrum beta-lactamase-
799 producing *Klebsiella pneumoniae* strains in a university hospital in Tunis,
800 Tunisia, 1999&-2005. *Clin. Microbiol. Infect.* **16**, 157–164 (2010).
801 28. Chen, L. *et al.* Carbapenemase-producing *Klebsiella pneumoniae*: molecular
802 and genetic decoding. *Trends Microbiol.* **22**, 686–696 (2014).

803 29. Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. & Brisse, S. Multilocus
804 sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin*
805 *Microbiol.* **43**, 4178–4182 (2005).
806 30. Wyres, K. L. & Holt, K. E. *Klebsiella pneumoniae* Population Genomics and
807 Antimicrobial-Resistant Clones. *Trend Microbiol* **24**, 944–956 (2016).
808 31. Long, S. W. *et al.* Population Genomic Analysis of 1,777 Extended-Spectrum
809 Beta-Lactamase-Producing *Klebsiella pneumoniae* Isolates, Houston, Texas:
810 Unexpected Abundance of Clonal Group 307. *MBio* **8**, e00489--17 (2017).
811 32. Potter, R. F. *et al.* Population Structure, Antibiotic Resistance, and
812 Uropathogenicity of *Klebsiella variicola*. *MBio* **9**, e02481-18 (2018).
813 33. David, S. *et al.* Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in
814 Europe is driven by nosocomial spread. *Nat. Microbiol.* (2019).
815 doi:10.1038/s41564-019-0492-8
816 34. Lam, M. M. C. *et al.* Genetic diversity, mobilisation and spread of the
817 yersiniabactin-encoding mobile element ICE*Kp* in *Klebsiella pneumoniae*
818 populations. *Microb Genom* **Jul 9**, (2018).
819 35. Lam, M. C. C. *et al.* Tracking key virulence loci encoding aerobactin and
820 salmochelin siderophore synthesis in *Klebsiella pneumoniae*. *Genome Med.* **10**,
821 77 (2018).
822 36. Wick, R. R., Heinz, E., Holt, K. E. & Wyres, K. L. Kaptive Web: user-friendly
823 capsule and lipopolysaccharide serotype prediction for *Klebsiella* genomes. *J*
824 *Clin Microbiol.* **56**, e00197-18 (2018).
825 37. Martínez-Romero, E. *et al.* Genome misclassification of *Klebsiella variicola*
826 and Klebsiella quasipneumoniae isolated from plants, animals and humans.
827 *Salud Publica Mex* **60**, 52–62 (2018).
828 38. Rodrigues, C. *et al.* Description of *Klebsiella africanensis* sp. nov., *Klebsiella*

829 *variicola* subsp. *tropicalensis* subsp. nov. and *Klebsiella variicola* subsp.
830 *variicola* subsp. nov. *Res. Microbiol.* **S0923-2508**, 30019–1 (2019).
831 39. Long, S. W. *et al.* Whole-genome sequencing of a human clinical isolate of the
832 novel species *Klebsiella quasivariicola* sp. nov. *Genome Announc.* **5**, e01057-
833 17 (2017).
834 40. Ondov, B. D. *et al.* Mash: Fast genome and metagenome distance estimation
835 using MinHash. *Genome Biol.* **17**, 132 (2016).
836 41. Fajardo-Lubia´n, A., Ben Zakour, N. L., Agyekum, A., Qi, Q. & Iredell, J. R.
837 Host adaptation and convergent evolution increases antibiotic resistance
838 without loss of virulence in a major human pathogen. *PLoS Pathog.* **15**,
839 e1007218 (2019).
840 42. Wong, J. L. C. *et al.* OmpK36-mediated Carbapenem resistance attenuates
841 ST258 *Klebsiella pneumoniae in vivo*. *Nat. Commun.* **10**, 3957 (2019).
842 43. Hauck, C. *et al.* Spectrum of excess mortality due to carbapenem-resistant
843 *Klebsiella pneumoniae* infections. *Clin. Microbiol. Infect.* **22**, 513–519 (2016).
844 44. Opoku-Temeng, C., Kobayashi, S. D. & DeLeo, F. R. *Klebsiella pneumoniae*
845 capsule polysaccharide as a target for therapeutics and vaccines. *Comput.*
846 *Struct. Biotechnol. J.* **17**, 1360–1366 (2019).
847 45. Venturini, C. *et al.* Fine capsule variation affects bacteriophage susceptibility
848 in *Klebsiella pneumoniae* ST258. *FASEB J.* **34**, 10801–10817 (2020).
849 46. Arena, F. *et al.* Population structure of KPC carbapenemase-producing
850 *Klebsiella pneumoniae* in a long-term acute-care rehabilitation facility:
851 identification of a new lineage of clonal group 101, associated with local
852 hyperendemicity. *Microb. genomics* **6**, e000308 (2020).
853 47. Ferrari, C. *et al.* Multiple *Klebsiella pneumoniae* KPC Clones Contribute to an
854 Extended Hospital Outbreak. *Front. Microbiol.* **10**, 2767 (2019).

855 48. Magiorakos, A.-P. *et al.* Multidrug-resistant, extensively drug-resistant and
856 pandrug-resistant bacteria: an international expert proposal for interim standard
857 definitions for acquired resistance. *Clin. Microbiol. Infect.* **18**, 268–281 (2012).
858 49. Wyres, K. L. *et al.* Genomic surveillance for hypervirulence and multi-drug
859 resistance in invasive *Klebsiella pneumoniae* from South and Southeast Asia.
860 *Genome Med.* **12**, 11 (2020).
861 50. Heinz, E., Brindle, R., Morgan-McCalla, A., Peters, K. & Thomson, N. R.
862 Caribbean multi-centre study of *Klebsiella pneumoniae*: whole genome
863 sequencing, antimicrobial resistance and virulence factors. *bioRxiv* (2019).
864 doi:10.1101/541136
865 51. Musicha, P. *et al.* Genomic analysis of *Klebsiella pneumoniae* isolates from
866 Malawi reveals acquisition of multiple ESBL determinants across diverse
867 lineages. *J. Antimicrob. Chemother.* **74**, 1223–1232 (2019).
868 52. Alvarez-Uria, G., Gandra, S., Mandal, S. & Laxminarayan, R. Global forecast
869 of antimicrobial resistance in invasive isolates of Escherichia coli and
870 *Klebsiella pneumoniae*. *Int. J. Infect. Dis.* **68**, 50–53 (2018).
871 53. Brolund, A. *et al.* Worsening epidemiological situation of carbapenemase-
872 producing Enterobacteriaceae in Europe, assessment by national experts from
873 37 countries, July 2018. *Eurosurveillance* **24**, (2019).
874 54. Bell, J. *et al. Gram-negative Sepsis Outcome Program 2019 Report*. (2019).
875 55. Yu, W.-L., Lee, M.-F., Tang, H.-J., Chang, M.-C. & Chuang, Y.-C. Low
876 prevalence of *rmpA* and high tendency of *rmpA* mutation correspond to low
877 virulence of extended spectrum β-lactamase-producing *Klebsiella pneumoniae*
878 isolates. *Virulence* **6**, 162–172 (2015).
879 56. Chen, L. & Kreiswirth, B. N. Convergence of carbapenem-resistance and
880 hypervirulence in *Klebsiella pneumoniae*. *Lancet Infect Dis* **18**, 2–3 (2018).

881 57. Lam, M. M. C. *et al.* Convergence of virulence and multidrug resistance in a
882 single plasmid vector in multidrug-resistant *Klebsiella pneumoniae* ST15. *J*
883 *Antimicrob Chemother.* (2019). doi:https://doi.org/10.1093/jac/dkz028
884 58. Chen, Y. *et al.* Preterm infants harbour diverse *Klebsiella* populations,
885 including atypical species that encode and produce an array of antimicrobial
886 resistance- and virulence-associated factors. *Microb. Genomics* **6**, (2020).
887 59. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in
888 caesarean-section birth. *Nature* **574**, 117–121 (2019).
889 60. Brisse, S. *et al. wzi* Gene Sequencing, a Rapid Method for Determination of
890 Capsular Type for *Klebsiella* Strains. *J. Clin. Microbiol.* **51**, 4073–4078
891 (2013).
892 61. Wick, R. R., Judd, L. M., Gorrie, C. & Holt, K. E. Unicycler: Resolving
893 bacterial genome assemblies from short and long sequencing reads. *PLoS*
894 *Comput Biol* **13**, e1005595 (2017).
895 62. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a
896 new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
897 63. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated
898 binning algorithm to recover genomes from multiple metagenomic datasets.
899 *Bioinformatics* **32**, 605–607 (2016).
900 64. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with
901 Kraken 2. *Genome Biol.* **20**, 257 (2019).
902 65. Lu, J., Bretwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating
903 species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, (2017).
904 66. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and
905 Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
906 67. Bialek-davenet, S. *et al.* Genomic Definition of Hypervirulent and Multidrug-

907 Resistant *Klebsiella pneumoniae* Clonal Groups. *Emerg. Infect. Dis.* **20**, 1812–
908 1820 (2014).
909 68. Neubauer, S. *et al.* A Genotype-Phenotype Correlation Study of SHV β-
910 Lactamases Offers New Insight into SHV Resistance Profiles. *Antimicrob.*
911 *Agents Chemother.* **64**, e02293-19 (2020).
912 69. Kidd, T. J. *et al.* Molecular mechanisms and virulence of colistin-resistant
913 *Klebsiella pneumoniae*. *Eur. Respir. J.* **48**, PA2625 (2016).
914 70. Cannatelli, A. *et al.* In vivo evolution to colistin resistance by PmrB sensor
915 kinase mutation in KPC-producing *Klebsiella pneumoniae* is associated with
916 low-dosage colistin treatment. *Antimicrob. Agents Chemother.* **58**, 4399–4403
917 (2014).
918 71. Cannatelli, A. *et al.* MgrB Inactivation Is a Common Mechanism of Colistin
919 Resistance in KPC-Producing *Klebsiella pneumoniae* of Clinical Origin.
920 *Antimicrob. Agents Chemother.* **58**, 5696 LP – 5703 (2014).
921 72. Drlica, K. & Zhao, X. DNA gyrase, topoisomerase IV, and the 4-quinolones.
922 *Microbiol. Mol. Biol. Rev.* **61**, 377–392 (1997).
923
924
925 **Author Contributions**
926 Study design: K.E.H. Data analysis: M.M.C.L., K.L.W. and K.E.H. Code
927 development: R.R.W., K.E.H., S.C.W., L.T.C., and K.L.W. Manuscript writing:
928 M.M.C.L., K.L.W., and K.E.H. All authors contributed to manuscript editing.
929
930 **Competing Interests**
931 None declared.

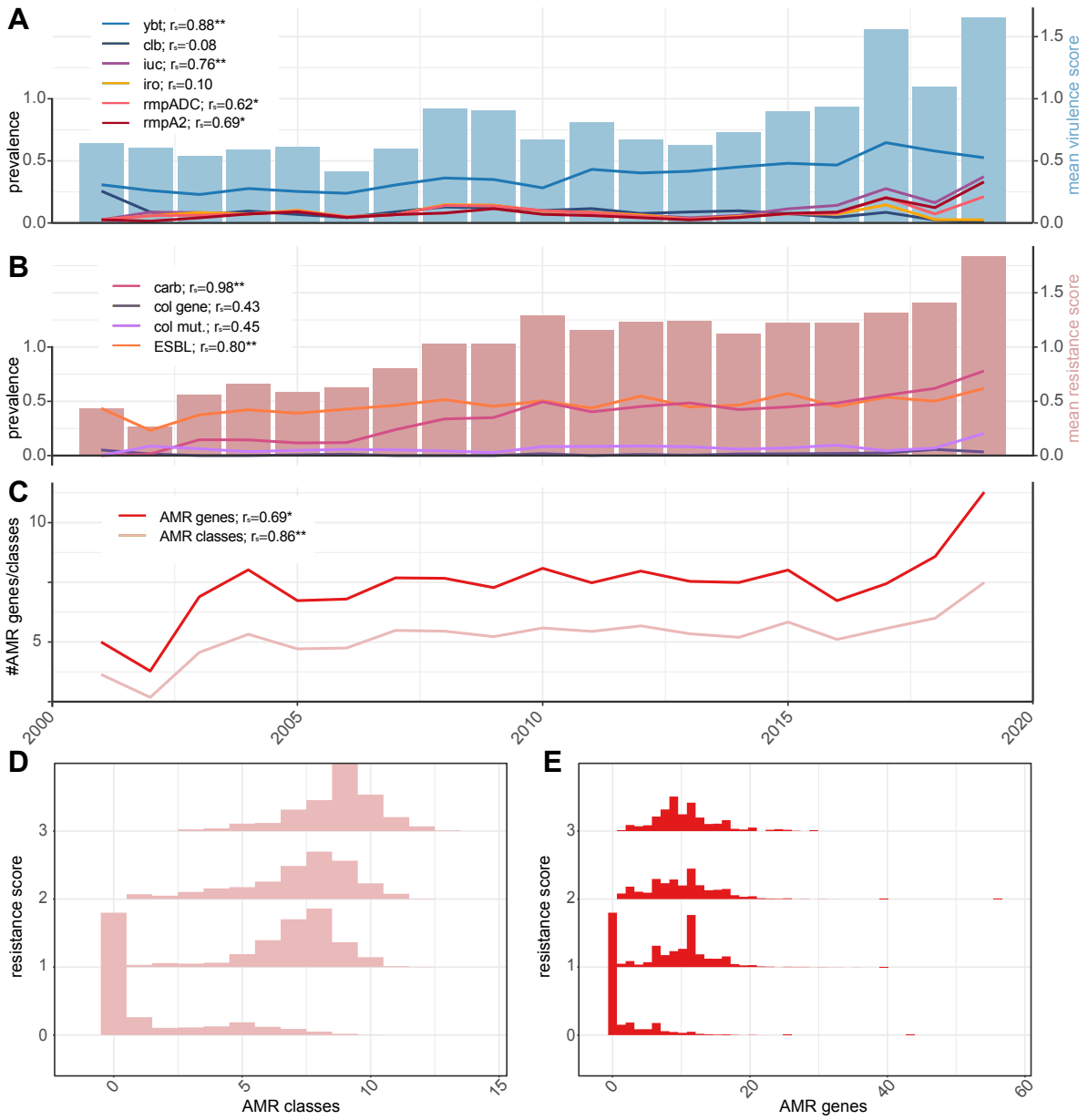

Figure 1. Relationships between Kleborate virulence and resistance scores and the prevalence of key virulence and antimicrobial resistance (AMR). Data shown summarise Kleborate results for non-redundant set of 9,705 publicly available *K. pneumoniae* genomes (**Table S2**). **(A)** Mean virulence score (barplot, right y-axis) and prevalence of individual virulence loci (line plots, left y-axis) over time. Ybt, yersiniabactin; clb, colibactin; iuc, aerobactin; iro, salmochelin; rmpADC, hypermucoidy *rmp* locus; rmpA2, *rmpA2* gene. Correlations between mean virulence score and prevalence of each locus are noted. **(B)** Mean resistance score (barplot, right y-axis) and prevalence of carbapenemases (carb), acquired colistin resistance genes (col gene), mutations in MgrB/PmrB (col mut) and genes conferring resistance to extended-spectrum β-lactams (ESBL) (line plots, left y-axis).

Correlations between mean resistance score and prevalence of each resistance type are noted. **(C)** Mean number of acquired AMR genes and classes, over time. Correlations with mean resistance score are noted. **(D)** Histograms showing total number of acquired AMR classes predicted per genome, stratified by resistance score. **(E)** Histograms showing total number of acquired AMR genes detected per genome, stratified by resistance score. Correlations reported in **A-C** are Spearman rank correlations; significance levels are indicated with asterisks: *$p<0.01$, **$p<0.001$.

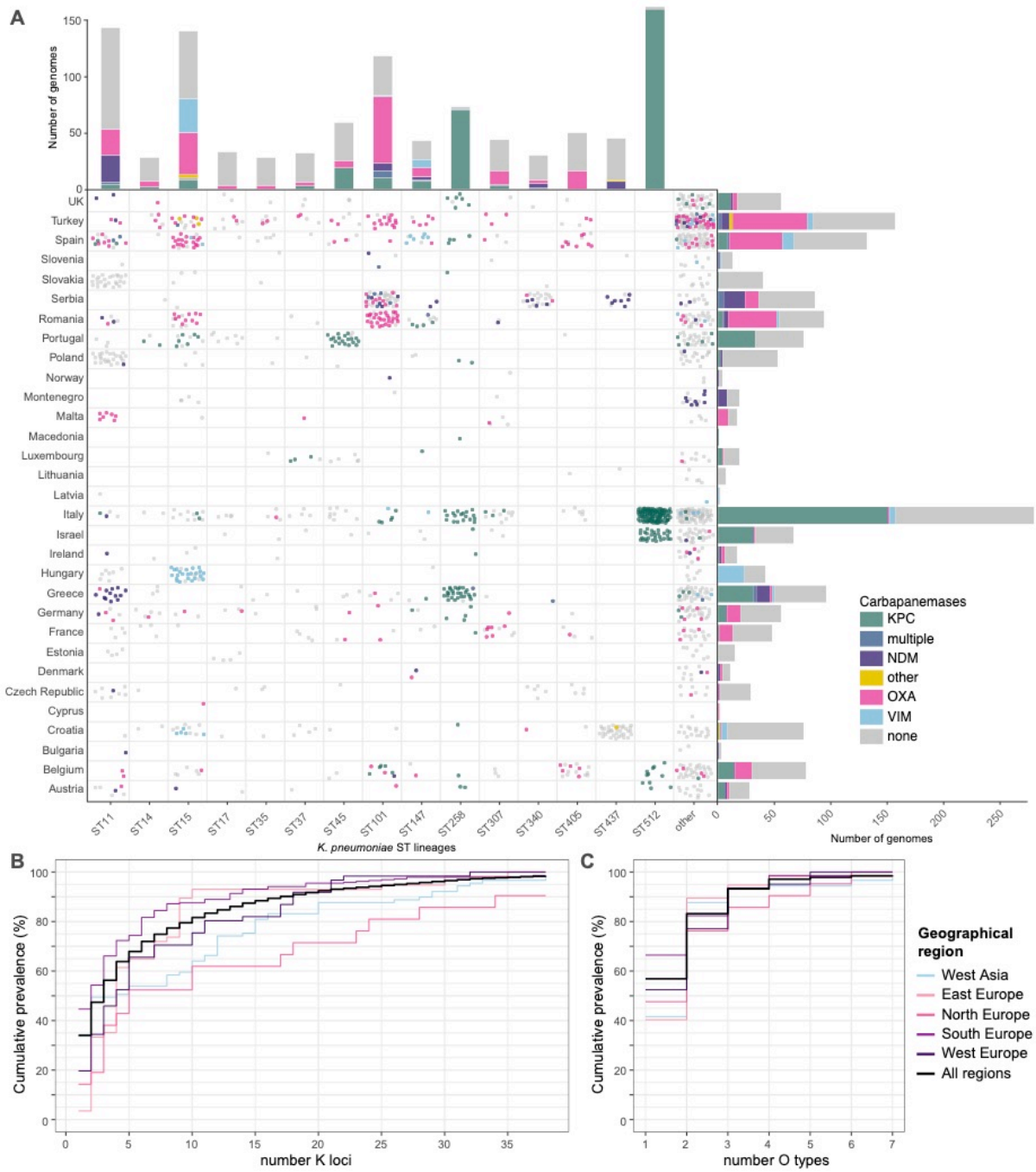

Figure 2. Kleborate genotyping results for European *K. pneumoniae* surveillance isolates. Data shown summarise Kleborate results for 927 carbapenem-non-susceptible and 697 carbapenem-susceptible *K. pneumoniae* genomes from the EuSCAPE study (data included in **Table S2**). **(A)** Geographical and lineage distribution of carbapenemase genes. Each circle represents a genome, colored by carbapenemase (see inset legend). Barplots summarise the number of genomes from each *K. pneumoniae* lineage (top) and country (right), colored by carbapenemase. (**B-C**) Cumulative prevalence of **(B)** capsule (K) locus

and **(C)** O antigen locus types, for carbapenem non-susceptible (meropenem MIC>2) isolates, ordered by overall prevalence. Thick line indicates curve for whole data set; others give results separately for different United Nations geographical regions (see inset legend).

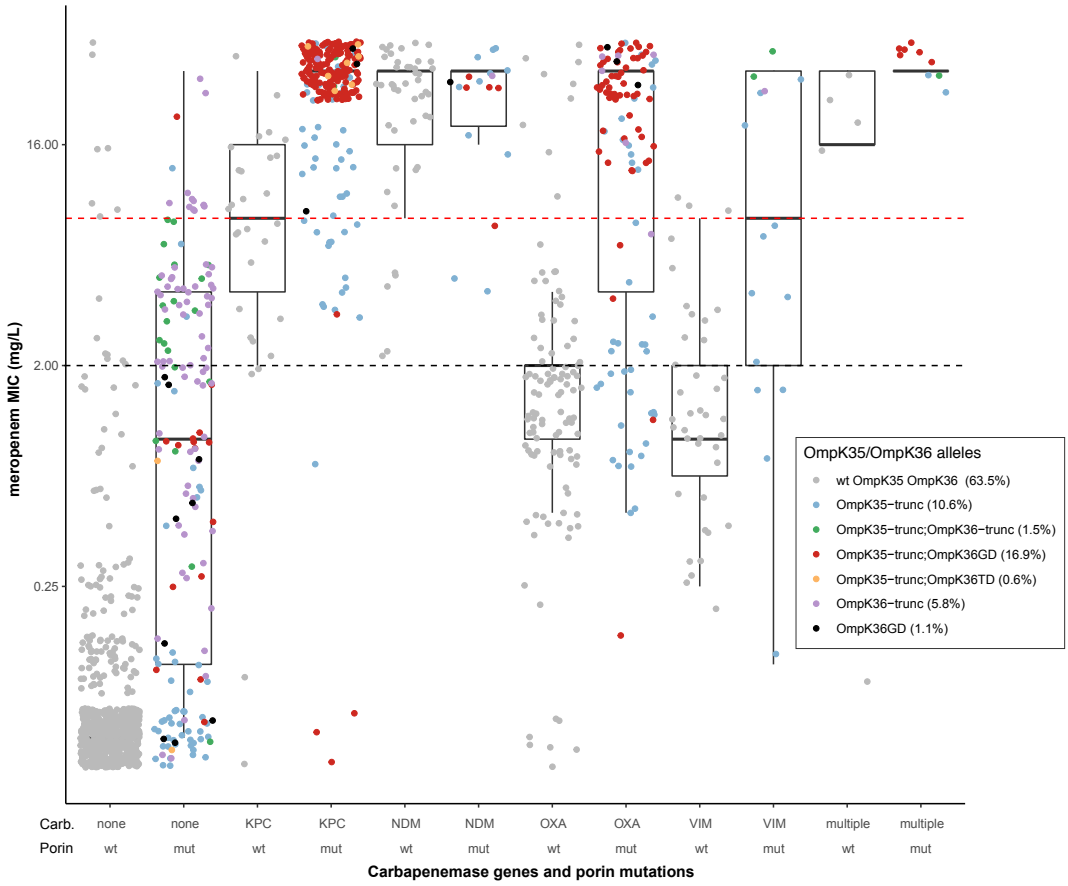

Figure 3. Distribution of meropenem MIC, stratified by Kleborate-detected carbapenemase genes and OmpK35/36 porin mutations, for European *K. pneumoniae* surveillance isolates. Data shown summarise Kleborate results for 1,490 *K. pneumoniae* genomes from the EuSCAPE study (data included in **Table S2**). Each circle represents the reported meropenem MIC for an isolate, coloured by type of porin mutation/s identified by Kleborate from the corresponding genome assembly (colour key in inset legend, prevalence of each genotype across 1490 genomes is indicated in brackets). Isolates are stratified by carbapenemase gene (enzymes labelled on x-axis) and OmpK mutations[41,42] reported by Kleborate. Wt, full-length OmpK35 and OmpK36 with no GD/TD insertion in the OmpK36 β-strand loop; mut, otherwise; trunc, truncation. Dashed lines indicate EUCAST breakpoints for clinical resistance (red, MIC >8) and non-susceptibility (black, MIC >2).

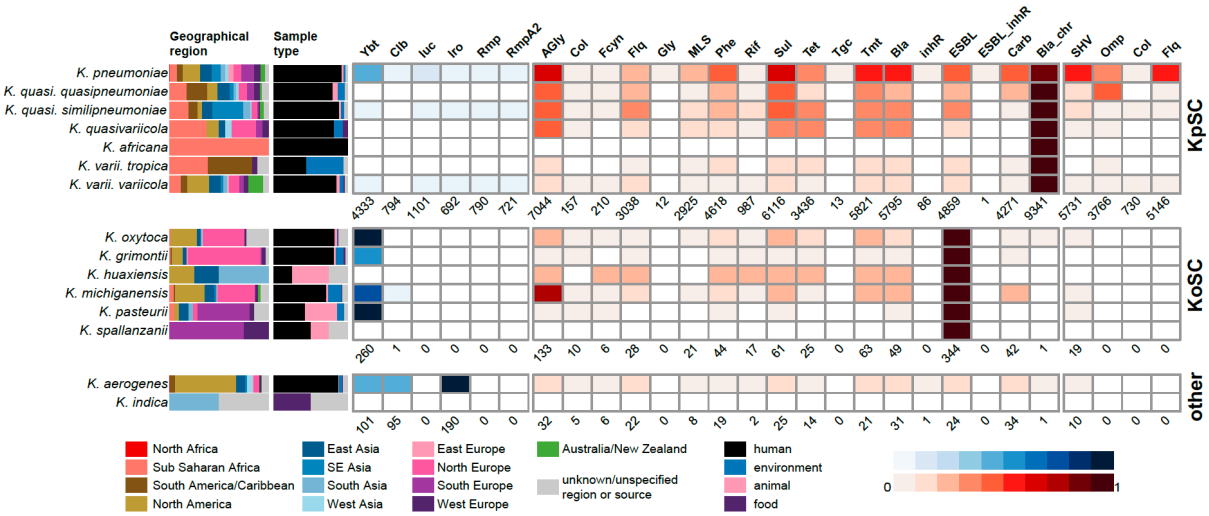

Figure 4. Summary of genome collection metadata, and Kleborate-derived virulence and antimicrobial resistance (AMR) genotypes, for all publicly available *Klebsiella* genomes. Data shown summarise Kleborate results for 11,277 non-redundant *Klebsiella* genomes publicly available as at July 17, 2020 (**Table S2**). From left to right: barplots showing source information by geographical region and sample type (coloured as per inset legend); heatmaps showing prevalence of virulence loci (blue) and predicted AMR drug classes (red) (as per inset scale bars). Genomes are summarised by species, ordered by species complex: KpSC, *K. pneumoniae* species complex; KoSC, *K. oxytoca* species complex; and other *Klebsiella*. In the heatmaps, the total number of genomes in which each type of virulence/AMR determinant was detected are indicated below each column. Column names are as follows: ybt, yersiniabactin; clb, colibactin; iuc, aerobactin; iro, salmochelin; rmp, hypermucoidy Rmp; rmpA2, hypermucoidy rmpA2; AGly, aminoglycosides; Col, colistin; Fcyn, fosfomycin; Flq, fluoroquinolone; Gly, glycopeptide; MLS, macrolides; Phe, phenicols; Rif, rifampin; Sul, sulfonamides; Tet, tetracyclines; Tgc, tigecycline; Tmt, trimethoprim; Bla, β-lactamases; inhR, β-lactamase inhibitor; ESBL, extended-spectrum β-lactamases; ESBL_inhR, extended-spectrum β-lactamase with resistance to β-lactamase inhibitors; Carb, carbapenemase; Bla_chr, intrinsic chromosomal β-lactamase; SHV,

mutations in SHV; Omp, truncations/mutations in *ompK35/ompK36*; Col, truncations in *mgrB/pmrB* conferring colistin resistance; Flq, mutations in *gyrA/parC* conferring resistance to fluoroquinolones.

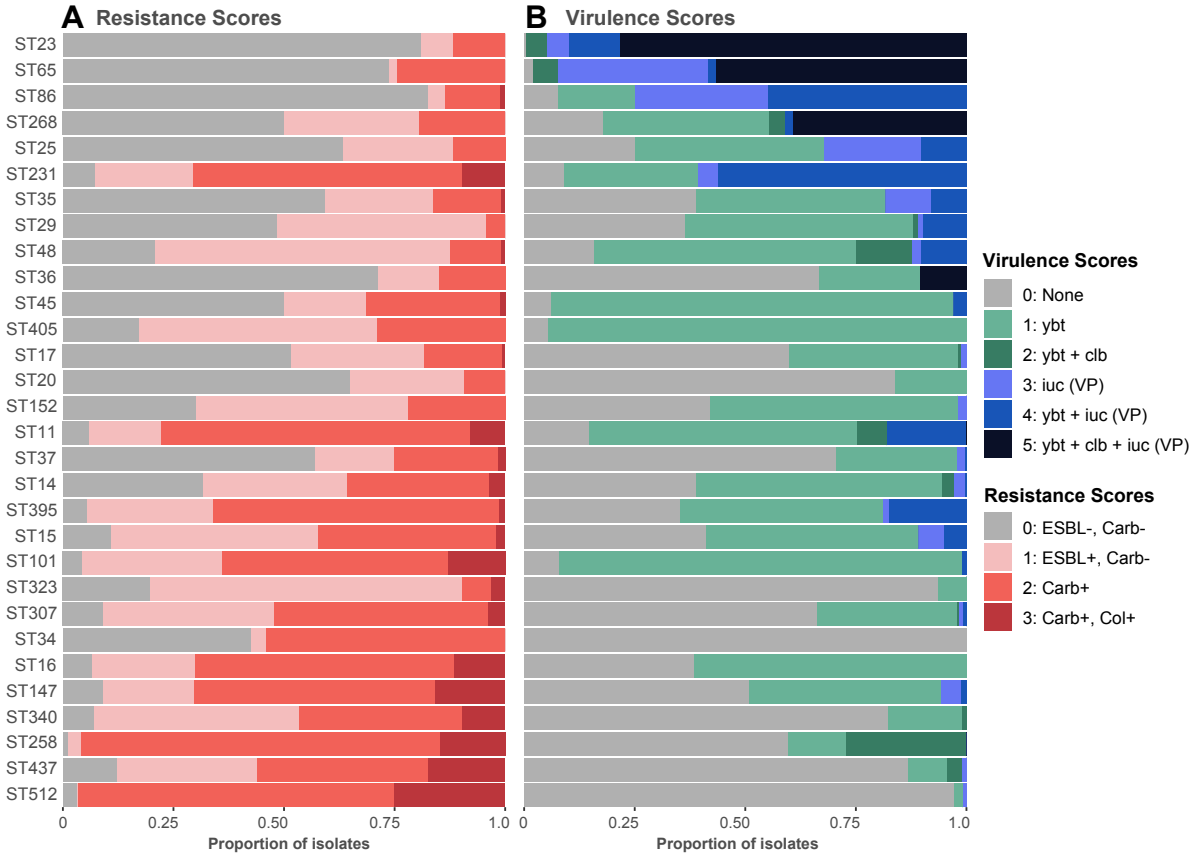

Figure 5. Distribution of resistance and virulence scores among genomes belonging to the 30 most common *K. pneumoniae* lineages. Data shown summarise Kleborate results for non-redundant set of 9,705 publicly available *K. pneumoniae* genomes (**Table S2**). Lineages were defined on the basis of multi-locus sequence types (STs) reported by Kleborate, and ordered from highest to lowest difference between mean virulence and mean resistance score. Minimum genome count per ST shown is 50. Ybt, yersiniabactin; clb, colibactin; iuc, aerobactin; VP, virulence plasmid; ESBL, extended-spectrum β-lactamase; Carb, carbapenemase; Col, colistin resistance determinant

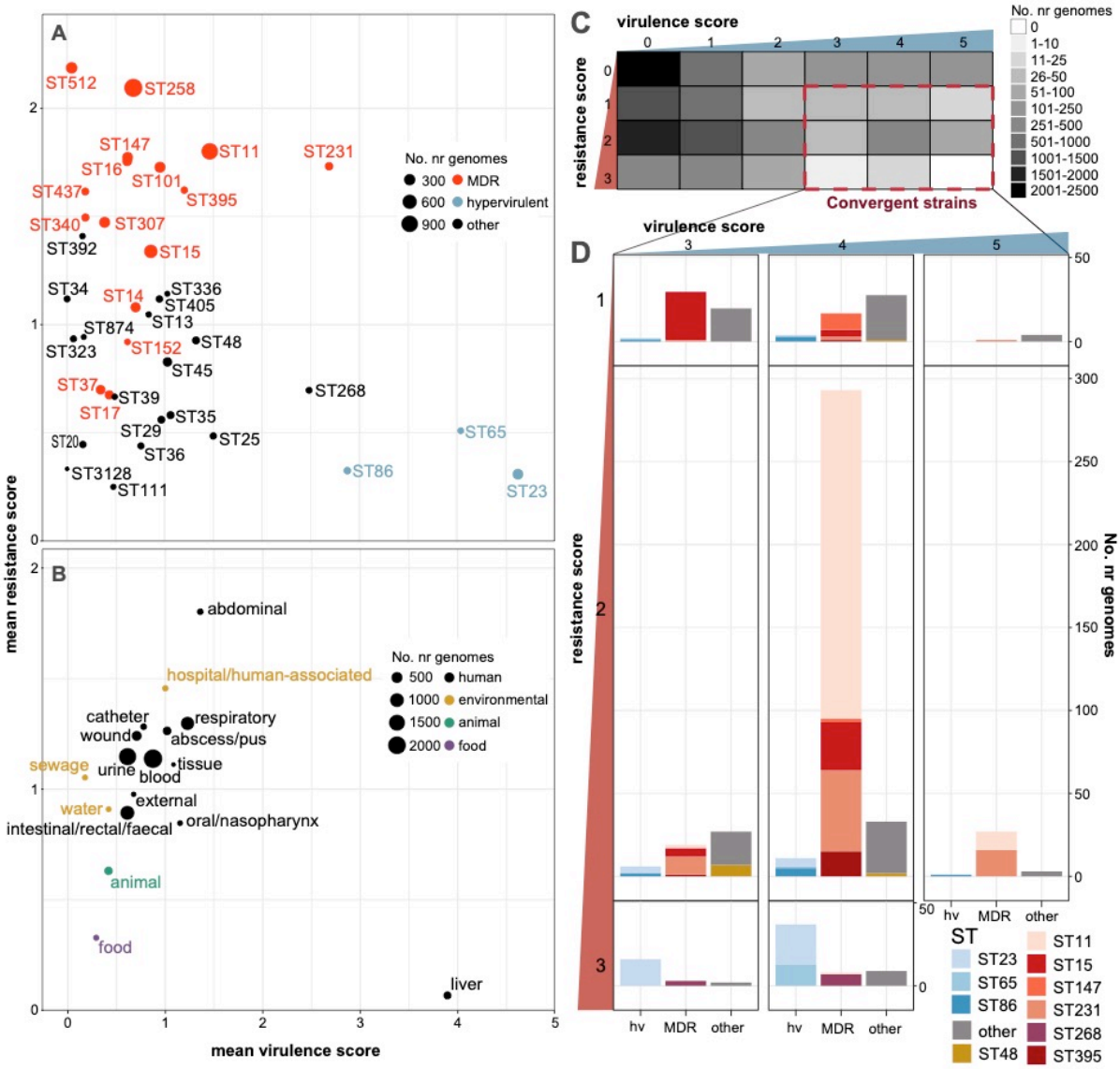

Figure 6. Insights from resistance and virulence scores. Data shown summarise Kleborate results for non-redundant set of 9,705 publicly available *K. pneumoniae* genomes (**Table S2**). (**A-B**) Mean resistance and virulence scores grouped by (**A**) lineage and (**B**) sample type. Each circle represents a single lineage (multi-locus sequence type, ST) or sample type as labelled; size indicates the number of genomes (as per inset legend); colour indicates groups per inset legend. **(C)** Heatmap showing number of genomes with each combination of resistance and virulence scores. Convergent genomes correspond to a virulence score ≥3 (carrying *iuc*) and resistance score of ≥1 (carrying ESBL and/or carbapenemase gene/s), as indicated by the red box. **(D)** Barplots showing lineage distribution of convergent genomes,

for each combination of resistance score and virulence score. Lineages are grouped into hypervirulent (hv), multidrug resistant (MDR) and other; and coloured by ST (as per inset legend).

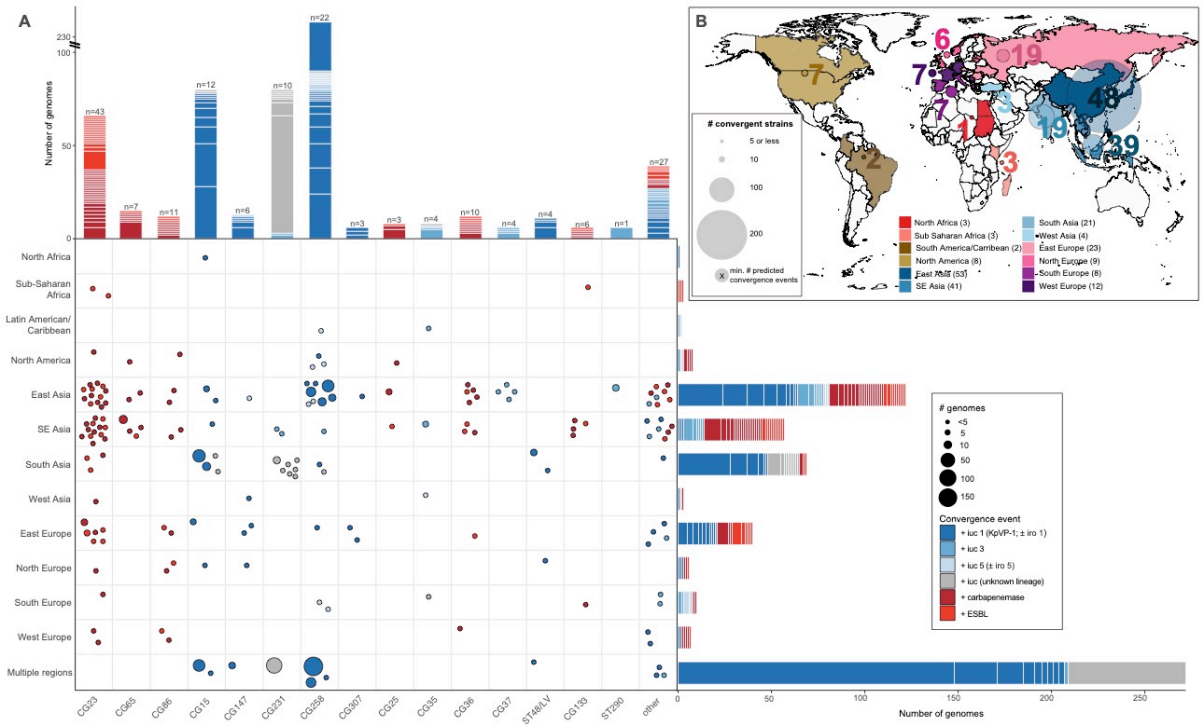

Figure 7. Convergence of AMR and virulence determinants in the *K. pneumoniae* population, identified by Kleborate analysis of public genomes. (A) Geographical and lineage distribution of convergence events. Each circle represents a unique convergence event (i.e. a monophyletic clade harbouring both ESBL/carbapenemase genes and *iuc*; see interactive tree at https://microreact.org/project/JDyan46yctyDh6weEUjWN, summary of events in **Table S8**, assignment of genomes to events in **Table S2**). Circles are scaled by the number of total genomes linked to the event and colored to indicate whether convergence is inferred to have occurred via acquisition of AMR gene/s (ESBL or carbapenemase/s) by a hypervirulent lineage or via acquisition of an *iuc*-encoding plasmid by an AMR lineage, as per inset legend. Marginal barplots show the number of convergence events (color blocks) and genomes (block heights) associated with each lineage (top) or geographical region (right). Lineages were defined on the basis of multi-locus sequence types (STs), number of convergence events estimated for each is labelled at the top of each bar. **(B)** Distribution of convergent genomes by location. Countries from which convergent genomes were detected are colored on the map; circles represent the number of convergent genomes detected in each UN-defined geographical region (indicated by color, as per inset legend), scaled and labelled

with the minimum estimated number of unique convergence events specific to each region (excluding inter-regional convergence events). The total number of convergence events affecting each region, including region-specific and inter-regional convergence events, are given in brackets in the inset legend.

Table 1. Genome features reported by Kleborate

Feature	Description
Assembly quality	Contig count, N50, largest contig, ambiguous bases
Identification	Species[16], MLST[29,67] (if *K. pneumoniae* species complex)
Acquired virulence determinants	Presence, genotypes, associated MGEs, truncations • yersiniabactin[34], • colibactin[34], • aerobactin[35], • salmochelin[35], • hypermucoidy loci *rmpADC* and *rmpA2*
Virulence score	0=no yersinabactin, colibactin or aerobactin; 1=yersiniabactin only; 2=yersiniabactin and colibactin (or colibactin only); 3= aerobactin without yersiniabactin or colibactin; 4= aerobactin with yersiniabactin (no colibactin); 5=yersiniabactin, colibactin and aerobactin
Serotype prediction	*wzi* allele and associated K locus[60] (default), Full K and O locus typing via *Kaptive*[36] (optional)
AMR determinants (optional)	
Acquired genes	Total count, alleles grouped by drug class, truncations
Mutations in core genes	SHV beta-lactamase (ESBL or inhibitors)[68], OmpK35/OmpK36 osmoporins[41,42] (carbapenems), MgrB/PmrB[69–71] (colistin), GyrA/ParC[72] (fluoroquinolones)

Number of drug classes	Excludes penicillins since resistance is intrinsic
Resistance score	1=ESBL; 2=Carbapenemase; 3=Carbapenemase plus colistin resistance; 0 otherwise

Table 2. Prevalence of virulence loci, ESBL and carbapenemase genes in non-redundant *Klebsiella* genomes

Species Complex	**Species**	Total no. genomes	Virulence prevalence	ESBL prevalence	Carbapenemase prevalence
K. pneumoniae species complex	*K. pneumoniae*	9705	Ybt: 4309, 44% Clb: 794, 8% Iuc: 1090, 11% Iro: 683, 7% Rmp: 782, 8% RmpA2: 716, 7%	4634, 48%	4173, 43%
	K. quasipneumoniae subsp. *quasipneumoniae*	119	-	31, 26%	29, 24%
	K. quasipneumoniae subsp. *similipneumoniae*	363	Ybt: 8, 2% Iuc: 6, 2% Iro: 4, 1% Rmp: 4, 1% RmpA2: 3, 0.8%	138, 38%	32, 9%
	K. quasivariicola	16	-	3, 19%	-
	K. africana	1	-	-	-
	K. variicola subsp. *variicola*	498	Ybt: 15, 3% Iuc: 4, 0.8% Iro: 5, 1% Rmp: 4, 0.8% RmpA2: 2, 0.4%	52, 10%	36, 7%
	K. variicola subsp. *tropica*	18	-	2, 11%	1, 6%
K. oxytoca species complex	*K. oxytoca*	98	Ybt: 96, 98%	9*, 9%	6, 6%
	K. grimontii	75	Ybt: 41, 55%	1*, 1%	3, 4%
	K. huaxiensis	4	-	1*, 25%	-
	K. michiganensis	144	Ybt: 102, 71% Clb: 1, 0.7%	21*, 15%	33, 23%
	K. pasteurii	21	Ybt: 21, 100%	1*, 5%	-

	K. spallanzanii	4	-	-*	-
NA	*K. aerogenes*	209	Ybt: 101, 48% Clb: 95, 45% Iro: 190, 91%	24, 11%	34, 16%
	K. indica	2	-	-	-

*excluding OXY genes that are conserved in *K. oxytoca* species complex