





# A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex

Margaret M. C. Lam <sup>1</sup>✉, Ryan R. Wick <sup>1</sup>, Stephen C. Watts<sup>2</sup>, Louise T. Cerdeira<sup>1</sup>, Kelly L. Wyres <sup>1</sup> & Kathryn E. Holt <sup>1,3</sup>

*Klebsiella 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.

<sup>1</sup>Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, VIC, Australia. <sup>2</sup>Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, VIC, Australia. <sup>3</sup>London School of Hygiene & Tropical Medicine, London, UK. ✉email: [margaret.lam@monash.edu](mailto:margaret.lam@monash.edu)

**K**lebsiella pneumoniae bacteria commonly colonize the mammalian gut, but are also recognized as a major public health threat due to their ability to cause severe infections in healthcare settings and their association with antimicrobial resistance (AMR)<sup>1,2</sup>. Reports of *K. pneumoniae* gut colonization frequencies vary by country and demographics; prevalence rates as low as 4% and as high as 87% have been reported<sup>3–6</sup>. *K. pneumoniae* colonization is implicated in chronic diseases of the gastrointestinal tract including inflammatory bowel disease and colorectal cancer<sup>7</sup>. There is also a growing body of evidence highlighting colonization as a reservoir for extraintestinal infections (urinary tract infection, pneumonia, wound or surgical site infections, sepsis) in vulnerable individuals such as neonates, the elderly, immunocompromized and hospitalized patients<sup>8</sup>. Treatment of healthcare-associated (HA) *K. pneumoniae* infections is often limited by multidrug resistance (MDR) resulting from the accumulation of horizontally acquired AMR genes and mutations in core genes<sup>2</sup>. Treatment is further complicated by increasing frequencies of strains producing extended-spectrum  $\beta$ -lactamases (ESBL) and/or carbapenemases, prompting increased reliance on colistin and  $\beta$ -lactam/ $\beta$ -lactamase inhibitor combinations<sup>9,10</sup>. The World Health Organization has accordingly prioritized *K. pneumoniae* as a target for new drugs and therapies<sup>11</sup>.

Outside healthcare settings, *K. pneumoniae* is also recognized as a causative agent of community-acquired infections including urinary tract infection and pneumonia, but also invasive infections such as pyogenic liver abscess, endophthalmitis and meningitis<sup>12,13</sup>. Invasive community-acquired infections are generally associated with so-called hypervirulent *K. pneumoniae* (hvKp) and are most commonly reported in East and Southeast Asia, or in individuals with East Asian ancestry<sup>12</sup>. Features associated with hvKp include a K1, K2 or K5 polysaccharide capsule and horizontally acquired virulence factors encoding the siderophores aerobactin (*Iuc*) and salmochelin (*Iro*), the genotoxin colibactin (*Cib*), and a hypermucooid phenotype (conferred by the *rmpADC* locus)<sup>14–18</sup>. HvKp are rarely MDR and most strains remain susceptible to drugs except ampicillin, to which *K. pneumoniae* are intrinsically resistant due to the chromosomally encoded  $\beta$ -lactamase SHV<sup>19</sup>. However, there have been increasing reports of hvKp carrying AMR plasmids and co-occurrence of AMR and virulence determinants in non-hvKp isolates. The convergence of AMR and virulence in *K. pneumoniae* potentiates invasive and difficult-to-treat infections, and at least one fatal outbreak has been documented in China where carbapenemase-producing hvKp are increasingly common<sup>20–24</sup>.

Research conducted in the pre-genomic era characterized 77 distinct capsular (K) serotypes<sup>25</sup>, nine O types<sup>26</sup> and variable AMR profiles amongst the *K. pneumoniae* population<sup>27,28</sup>, indicating a diverse genetic and phenotypic landscape<sup>15,29</sup>. In recent years, genomic studies have provided key insights into the population structure of *K. pneumoniae* (recently summarized in Wyres et al.<sup>16</sup>), revealing hundreds of deep-branching phylogenetic lineages comprising sequence types (STs) or clonal groups (CGs) defined by the seven-gene multi-locus sequence typing (MLST) scheme<sup>29</sup>. Some of these correspond to lineages (i.e. STs and CGs) that have accumulated large numbers of AMR genes that have become globally distributed (e.g. CG258, CG15, ST307); these are dubbed MDR clones and have been linked with HA infections and hospital outbreaks worldwide<sup>30</sup>. Others carry a high load of virulence genes (e.g. CG23, CG65, CG86) and are recognized as hvKp associated with community-acquired infections. Further distinguishing MDR from hvKp clones are their K and O antigen profiles, with the former displaying a diverse range of K and O biosynthesis loci as a result of homologous recombination between strains, while hvKp rarely deviate from the K1, K2 or K5 types<sup>16</sup>.

Importantly, genomic characterization of clinical isolates identified as *K. pneumoniae* via biochemical tests or mass spectrometry (MALDI-TOF) has revealed the existence of multiple related species and subspecies, which together form the *K. pneumoniae* species complex (KpSC). These differ by 3–4% nucleotide divergence across core chromosomal genes, but share the same pool of AMR and virulence genes<sup>16</sup>. Infections and outbreaks caused by other KpSC members have been reported but they generally account for a significantly lower disease burden than *K. pneumoniae* (10–20%)<sup>19,31,32</sup>. Genomics has also clarified that the two *K. pneumoniae* subspecies originally defined by distinct and unusual disease manifestations (subsp. *rhinoscleromatis* which causes a progressive and chronic granulomatous infection known as rhinoscleroma, and subsp. *ozaenae* which causes atrophic rhinitis or ozena) actually represent CGs of *K. pneumoniae* (CG3 and CG90)<sup>15</sup>. Like hvKp clones, these strains also express specific capsule types (K3, K4 and K5) alongside aerobactin and another acquired siderophore, yersiniabactin (Ybt)<sup>16</sup>.

Due to its clinical importance and increasing AMR, *K. pneumoniae* is increasingly the focus of surveillance efforts and molecular epidemiology studies. The sheer volume of clinically relevant molecular targets renders whole-genome sequencing (WGS) the most cost-efficient characterization approach, however extracting and interpreting clinically important features is challenging. To address this, we have developed Kleborate, a genotyping tool designed specifically for *K. pneumoniae* and the associated species complex, which consolidates detection and genotyping of key virulence and AMR loci alongside species, lineage (ST) and predicted K and O antigen serotypes directly from genome assemblies. Here we describe Kleborate and demonstrate its utility by application to publicly available datasets. First, we show that Kleborate can rapidly recapitulate and augment the key findings from a recent large-scale European genomic surveillance study<sup>33</sup>. Next, we apply Kleborate to a curated collection of 13,156 publicly available WGS to further showcase its utility and derive novel insights into the global epidemiology of *Klebsiella* AMR, virulence and convergence. Finally, we show that Kleborate can also be applied to detect clinically relevant genotypes from metagenome-assembled genomes (MAGs).

## Results

**Integrated genomic framework and genotyping tool.** Our goal was to develop a single tool that can rapidly extract genotype information that is clinically relevant to *K. pneumoniae* and other members of the species complex in order to support genomic epidemiology and surveillance. We have previously reported genotyping schemes for the acquired *K. pneumoniae* virulence loci *ybt*, *clb*, *iuc* and *iro*<sup>34,35</sup> (whose detection and typing is implemented in early versions of Kleborate), and also K and O antigen typing implemented in the software Kaptive<sup>36</sup>. Here we expand the Kleborate framework to include additional features including taxonomic assignment to species and subspecies, assignment to lineages via seven-locus MLST, detection and genotyping of the *rmp* hypermucooid locus and the *rmpA2* gene, and identification of AMR determinants (mutations and horizontally acquired genes, including assignment of SHV  $\beta$ -lactamase alleles as either ESBL,  $\beta$ -lactamase inhibitor resistance, or intrinsic ampicillin resistance only, see “Methods” and Supplementary note 3). Kleborate can optionally call Kaptive for K/O antigen prediction.

Unlike generic AMR or virulence typing tools, we include only genetic features for which there is strong evidence of an associated phenotype in *K. pneumoniae* that has confirmed

**Table 1** Genome features reported by Kleborate.

Feature	Description
Assembly quality	Contig count, N50, largest contig, ambiguous bases
Identification	Species <sup>16</sup> , multi-locus sequence typing (MLST) <sup>29,85</sup> (if <i>K. pneumoniae</i> species complex)
Acquired virulence determinants	Presence, genotypes, associated mobile genetic elements (MGEs), truncations <ul style="list-style-type: none"> <li>• yersiniabactin<sup>34</sup>,</li> <li>• colibactin<sup>34</sup>,</li> <li>• aerobactin<sup>35</sup>,</li> <li>• salmochelin<sup>35</sup>,</li> <li>• hypermucoidy loci <i>rmpADC</i> and <i>rmpA2</i></li> </ul>
Virulence score	0 = no yersiniabactin, 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	<i>wzi</i> allele and associated K locus <sup>77</sup> (default), Full K and O locus typing via <i>Kaptive</i> <sup>36</sup> (optional)
AMR determinants (optional)	
Acquired genes	Total count, alleles grouped by drug class, truncations
Mutations in core genes	SHV beta-lactamase (extended-spectrum beta-lactamase/ESBL or inhibitors) <sup>86</sup> , OmpK35/OmpK36 osmoporins <sup>41,42</sup> (carbapenems), MgrB/PmrB <sup>57-59</sup> (colistin), GyrA/ParC <sup>76</sup> (fluoroquinolones)
Number of drug classes	Excludes penicillins since resistance is intrinsic
Resistance score	1 = ESBL; 2 = Carbapenemase; 3 = Carbapenemase plus colistin resistance; 0 otherwise

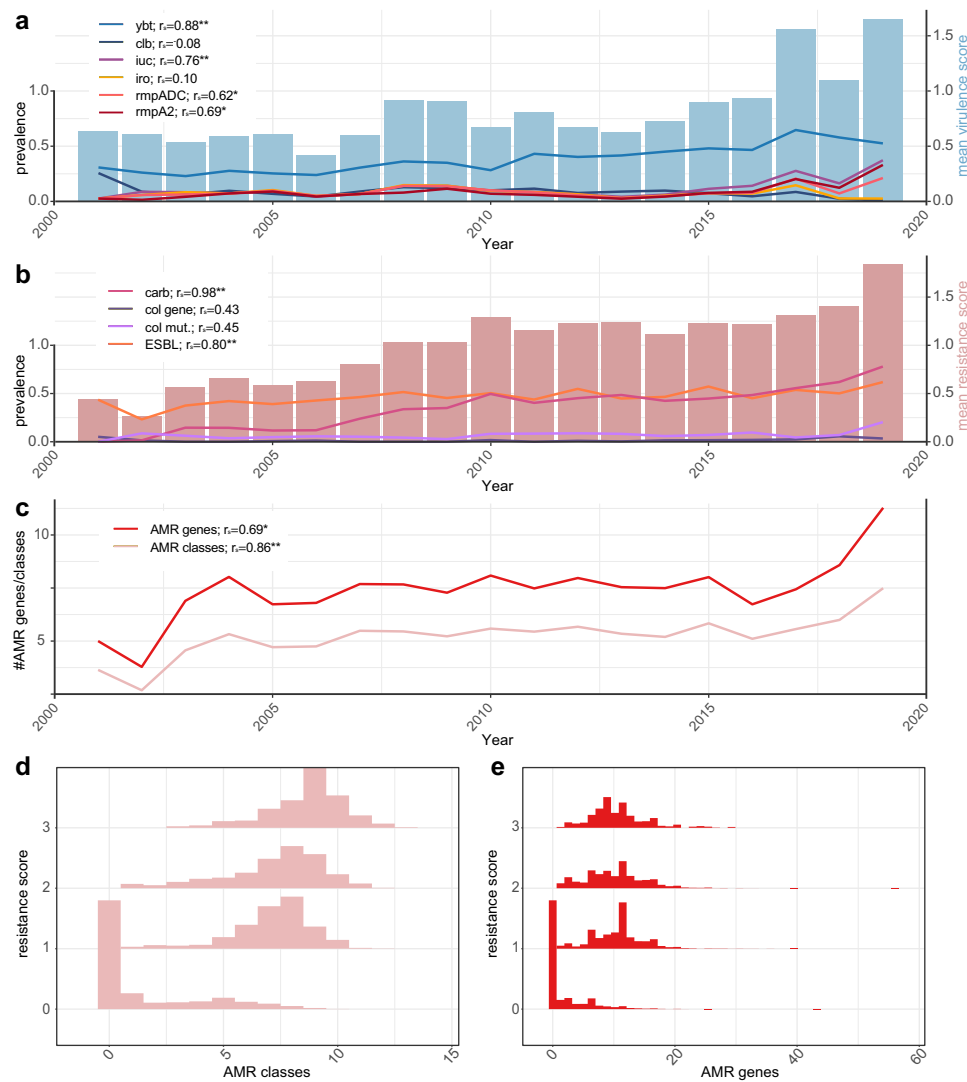
clinical relevance based on published experimental data (Table 1). These are reported in a manner that facilitates interpretation, including summarizing virulence and AMR genotypes into scores that reflect escalating clinical risk in *K. pneumoniae* infections. Kleborate features are summarized in Table 1 and methodological details for genotyping are provided in “Methods”. For a typical 5.5 Mbp genome, a Kleborate run including AMR typing takes <10 s on a laptop, while robust K and O serotype prediction using *Kaptive*<sup>36</sup> adds an additional ~1 min. Results are output in tab-delimited format, making it easy to integrate Kleborate into existing workflows.

**Species and subspecies assignment.** The taxonomy of *Klebsiella* is rapidly evolving, with several new species and subspecies recently identified<sup>37-39</sup>. As a consequence, many genomes in public databases are incorrectly assigned. We therefore introduced a custom approach for rapid and accurate species and subspecies identification for *Klebsiella*, based on Mash distances<sup>40</sup> to a taxonomically curated genome set (representative tree in Supplementary Fig. 1a, b), avoiding the need for users to download large reference genome databases (see “Methods”). This approach was validated using a set of  $n = 285$  diverse clinical isolates and compared with species assignments based on the read-based taxonomic classifier Kraken2 (details in Supplementary Note 1, Supplementary Data 1, Supplementary Fig. 1).

**Virulence and AMR scores.** Genomes are scored according to the clinical risk associated with the AMR and virulence loci that are detected (see “Methods”). Here we take advantage of the structured distribution of AMR and virulence determinants within the *K. pneumoniae* population<sup>14</sup> to summarize the genotyping data with simple numerical summary scores that reflect the accumulation of loci contributing to clinically relevant AMR or hypervirulence: virulence scores range from 0 to 5, depending on the presence of key loci associated with increasing risk (yersiniabactin < colibactin < aerobactin, see the detailed rationale in “Methods”); resistance scores range from 0 to 3, based on detection of genotypes warranting escalation of antimicrobial therapy (ESBL < carbapenemase < carbapenemase plus colistin resistance, see Table 1). These simple numerical scores facilitate downstream analyses including trend detection. For example, analysis of a non-redundant subset of 9,705 publicly available *K. pneumoniae* genomes (see below, Supplementary Data 2) showed increasing AMR and virulence scores over time

(barplots in Fig. 1a, b). The virulence and resistance scores were correlated not only with the prevalence of individual components that contribute to the scores, but also with other components that are co-distributed in the population (lines in Fig. 1a, b). For example, the frequencies of *rmpADC* and *rmpA2* loci over time were correlated with the virulence score (Fig. 1a); and the resistance score was correlated with the mean number of acquired AMR genes and associated drug classes (excluding ESBLs, carbapenemases and colistin which contribute to the score) (Fig. 1c). Consistent with this, genomes with resistance scores >0 (assigned based on the presence of ESBL and/or carbapenemase genes) typically carry many additional AMR genes conferring resistance to multiple drug classes (Fig. 1d, e). The distribution of virulence scores in this genome set differed by isolate source, and skewed higher in human clinical specimens (mean 0.86) compared to human gut carriage (mean 0.53,  $p < 1 \times 10^{-15}$ ), animal (mean 0.39,  $p < 1 \times 10^{-11}$ ), or environmental samples (mean 0.49,  $p < 1 \times 10^{-8}$ ) (see Supplementary Fig. 2). Isolates from liver abscess (associated with hypervirulent *K. pneumoniae*) had particularly high scores (mean 3.8, median 5), reflecting the presence of *iuc* as expected for this relatively rare infection type; in contrast other infection types most commonly have scores of 0 (40–60%) or 1 (reflecting presence of *ybt* without *iuc*; 20–40%), as expected for opportunistic *K. pneumoniae* infections. Notably, there were also animal and environmental isolates with non-zero virulence scores (5–25%), reflecting the cycling of *K. pneumoniae* between ecological niches<sup>2</sup>. Reducing the data to key axes of virulence and AMR also facilitates exploration of subpopulations associated with AMR, virulence or convergence of both traits; such as specific *K. pneumoniae* lineages or specimen types (see below). It is important to note though that the scores summarize the detection of specific genetic determinants, and are not direct predictions of clinical virulence or antibiotic resistance.

**Rapid genotyping of clinical isolates from a large-scale surveillance study.** We applied Kleborate to analyse all *K. pneumoniae* clinical isolate genomes deposited in RefSeq by the EuSCAPE surveillance study (927 carbapenem-non-susceptible, 697 carbapenem-susceptible; see Supplementary Data 2)<sup>33</sup>. Kleborate rapidly and accurately reproduced key findings from the original study, which were originally derived from multi-step analyses comprising five independent tools and four independent databases (each from a different public repository, one with

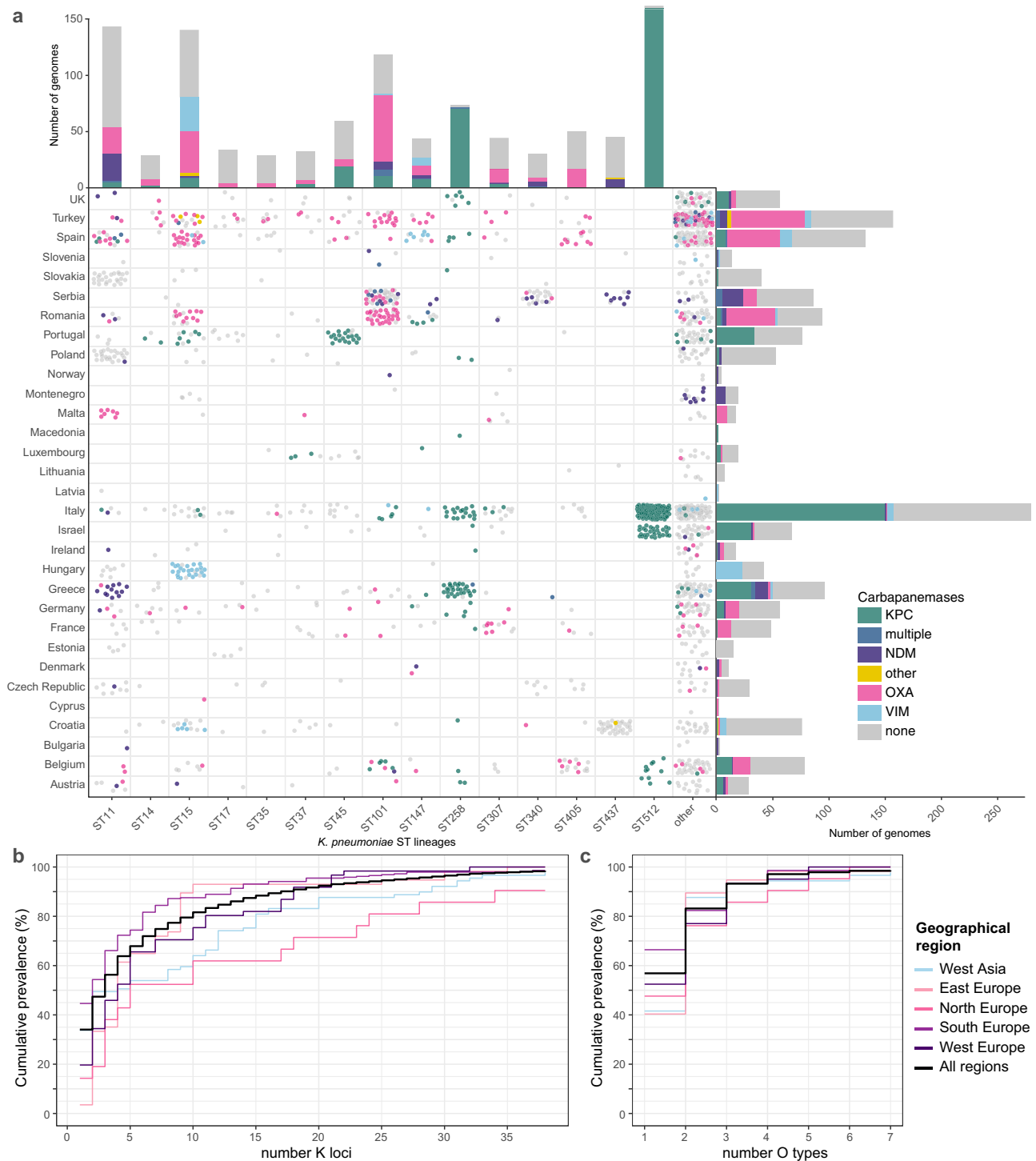


**Fig. 1 Relationships between Kleborate virulence and resistance scores and the prevalence of key virulence and antimicrobial resistance (AMR).** Data shown summarize Kleborate results for non-redundant set of 9705 publicly available *K. pneumoniae* genomes (Supplementary Data 2). **a** Barplot shows mean virulence score per year (right y-axis), line plots show the prevalence of individual virulence loci per year (left y-axis). Ybt yersiniabactin, clb colibactin, iuc aerobactin, iro salmochelin, rmpADC hypermucoidy *rmp* locus, rmpA2 *rmpA2* gene. Spearman correlation coefficients ( $r_s$ ) between the mean virulence score and prevalence of each locus are noted. *P* values from two-sided statistical testing: ybt =  $2.2 \times 10^{-16}$ , clb = 0.74, iuc = 0.0002, iro = 0.68, rmpADC = 0.005, rmpA2 = 0.002. **b** Barplot shows mean resistance score per year (right y-axis) and line plots show the prevalence of carbapenemases (carb), acquired colistin resistance genes (col gene), mutations in MgrB/PmrB (col mut) and genes conferring resistance to extended-spectrum  $\beta$ -lactams (ESBL) (left y-axis). Spearman correlation coefficients ( $r_s$ ) between mean resistance score and prevalence of each resistance type are noted. *P* values: carb =  $8.3 \times 10^{-6}$ , col gene = 0.06, col mut = 0.05, ESBL =  $5.56 \times 10^{-5}$ . **c** Mean number of acquired AMR genes and classes, over time. Spearman correlation coefficients with mean resistance score are noted. *P* values: AMR genes = 0.001, AMR classes =  $2.2 \times 10^{-16}$ . **d** Histograms showing total number of acquired AMR classes predicted per genome, stratified by resistance score. **e** Histograms showing a total number of acquired AMR genes detected per genome, stratified by resistance score. Spearman correlation coefficients are shown in a-c; significance levels are indicated with asterisks: \**p* < 0.01, \*\**p* < 0.001.

additional manual curation): (i) 70.2% of carbapenem-non-susceptible genomes ( $n = 651/927$ ) carried carbapenemases, mainly KPC-3, OXA-48, KPC-2 and NDM-1; (ii) these were dominated by a few major clones, ST11, ST15, ST45, ST101, ST258, and ST512; (iii) individual countries were associated with specific carbapenemase/clone combinations (Fig. 2a). A detailed comparison of the results reported by Kleborate versus those reported in the original study is provided in Supplementary Note 2 and Supplementary Data 3.

In addition to the detection of carbapenemase genes, Kleborate also identified porin defects, which are known to contribute to the carbapenem-resistance phenotype<sup>41,42</sup>, in 36.5% of EuSCAPE

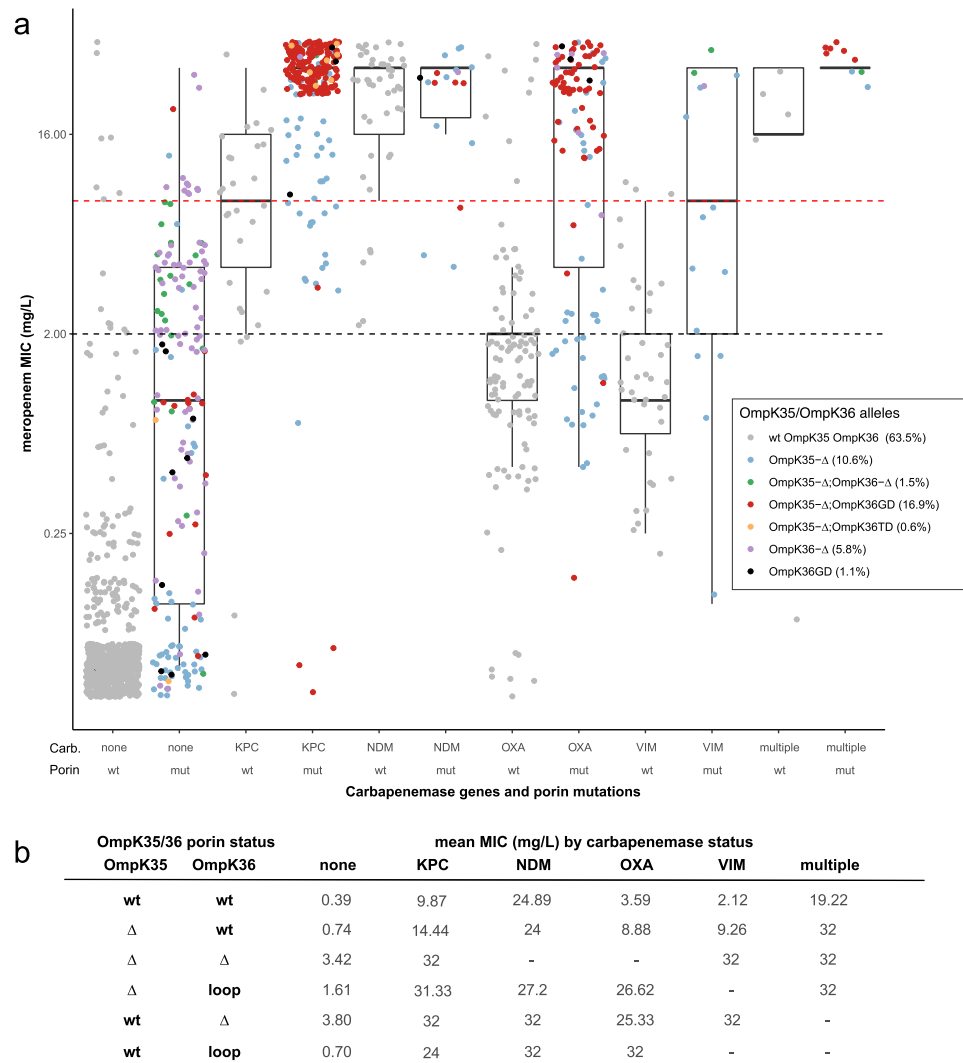
genomes (including 60% of those with carbapenemase genes and 19.9% of those without carbapenemase genes). These defects included truncation/deletion of OmpK35 and/or OmpK36 (also considered in the original study) as well as GD or TD insertions in the OmpK36  $\beta$ -strand loop<sup>41</sup> (these insertions were not analysed in the original study, but here were detected in 17.9% of genomes including 18 with no porin deletion). Figure 3 shows meropenem MICs stratified by combinations of porin defect and carbapenemase identified by Kleborate. OmpK mutations were associated with elevation of meropenem MIC, although in the absence of a carbapenemase gene the effect was only clinically significant (>2 mg/mL, the current EUCAST cut-off for insusceptibility) for loss



**Fig. 2 Kleborate genotyping results for European *K. pneumoniae* surveillance isolates.** Data shown summarize Kleborate results for 927 carbapenem-non-susceptible and 697 carbapenem-susceptible *K. pneumoniae* genomes from the EuSCAPE study (data included in Supplementary Data 2). **a** Geographical and lineage distribution of carbanemase genes. Each circle represents a genome, coloured by carbanemase (see inset legend). Barplots summarize the number of genomes from each *K. pneumoniae* lineage (top) and country (right), coloured by carbanemase. **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).

of OmpK36 (Fig. 3). Importantly, while carbapenemase genes were clearly associated with an elevation in meropenem MIC, there was variation depending on enzyme type. KPC and NDM were each associated mean MIC above the clinical cut-off for resistance (>8 mg/mL) even in isolates with wildtype OmpK35 and OmpK36; however OXA and VIM enzymes on average were

associated with MICs above the insusceptible cut-off (>2 mg/mL) in wildtype OmpK35/OmpK36 isolates, and required the presence of OmpK mutations to reach the resistance cut-off (>8 mg/mL) (Fig. 3). This highlights both (i) the importance of porin defects—including the OmpK36  $\beta$ -strand loop insertions—for full expression of carbapenem resistance in *K. pneumoniae*; and (ii) the

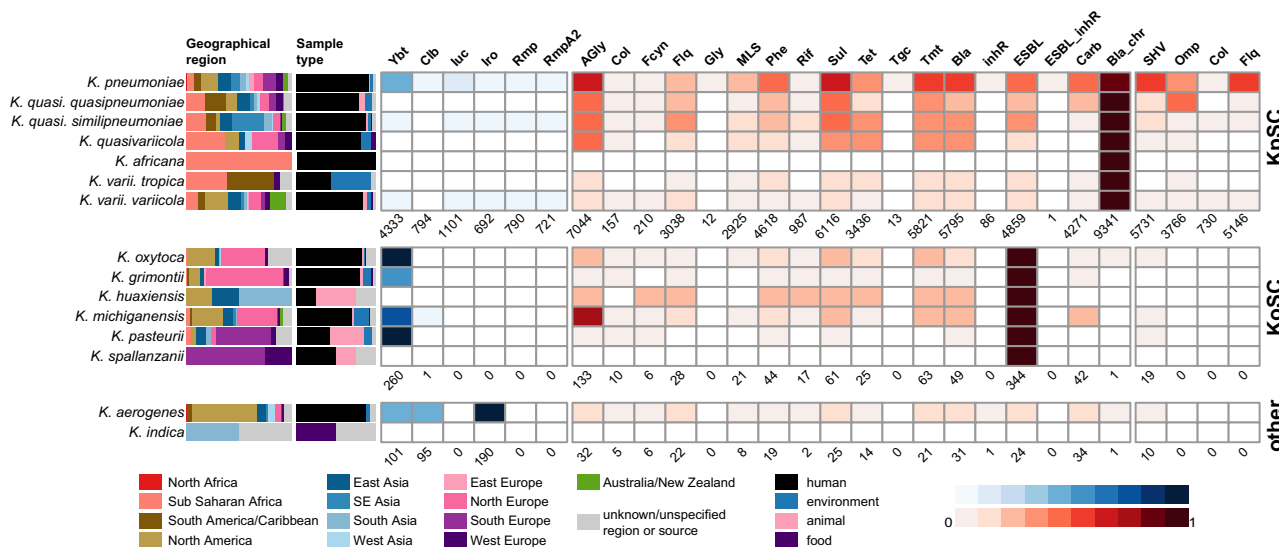


**Fig. 3** Distribution of meropenem MIC, stratified by Kleborate-detected carbapenemase genes and OmpK35/36 porin mutations, for European *K. pneumoniae* surveillance isolates. **a** Data shown summarize Kleborate results for 1490 *K. pneumoniae* genomes from the EuSCAPE study (data included in Supplementary Data 2). 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, the 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<sup>41,42</sup> reported by Kleborate. Wt, full-length OmpK35 and OmpK36 with no GD/TD insertion in the OmpK36  $\beta$ -strand loop; mut, otherwise;  $\Delta$ , missing/truncated. Dashed lines indicate EUCAST breakpoints for clinical resistance (red, MIC > 8) and non-susceptibility (black, MIC > 2). For each boxplot, the length of the box corresponds to the interquartile range with the centre line corresponding to the median, and the whiskers represent the minimum and maximum values. **b** Mean meropenem MIC values for the 1490 EuSCAPE isolates, grouped by the combination of porin gene status and the presence of carbapenemase genes. Porin status is expressed as:  $\Delta$ , missing/truncated; loop, GD or TD insertion in the OmpK36  $\beta$ -strand loop; wt, otherwise.

complex nature of some AMR mechanisms, which extends beyond mere presence or absence of a single acquired gene or mutation.

The rise in carbapenem-resistant *K. pneumoniae* infections in hospitals and its associated morbidity and mortality<sup>43</sup> has led to increased interest in alternative control strategies such as vaccines, phage therapy and antibody therapy, key targets for which are the K and O surface antigens<sup>44–47</sup>. Kleborate confidently identified K and O biosynthesis loci in 98.3% and 99.1% of EuSCAPE genomes, respectively, including 87 distinct K loci and 11 distinct O loci (Supplementary Figs. 3 and 4). Amongst carbapenem-non-susceptible isolates (meropenem MIC > 2), 38 distinct K types were identified and the most common were KL107 ( $n = 173$ ), KL17 ( $n = 67$ ), KL106 ( $n = 41$ ), KL24 ( $n = 35$ ), KL15 ( $n = 19$ ) and KL36 ( $n = 13$ ). Seven distinct O types were detected among these genomes, and the most common were O2 ( $n = 294$ ), O1 ( $n = 136$ ) and O4 ( $n = 52$ ). Overall, the data

suggest an intervention would need to be effective against six K types or two O types in order to provide coverage of 80% of carbapenem-resistant infections in Europe (Fig. 2b, c). However, it is important to explore the impact of population structure on these findings, specifically the impact of local clonal expansions. Kleborate aides this type of analysis by providing ST and other genotyping information alongside the K and O locus types, which can be viewed in the context of geographic information. Doing so revealed that each of the top three K loci was dominated by a single ST (83.5% of KL107 were ST512; 93.0% of KL105 were ST11; 91.4% of KL17 were ST101). Importantly, the vast majority of ST512-KL107 genomes (75.3%) originated from Italy where this ST is known to be locally circulating<sup>48,49</sup>, while 58% of ST11-KL105 originated from Poland and Slovakia, and 56% of ST101-KL17 originated from Serbia and Romania. When these putative local expansions were excluded, the top 6 K loci were (KL24,



**Fig. 4 Summary of genome collection metadata, and Kleborate-derived virulence and antimicrobial resistance (AMR) genotypes, for all publicly available *Klebsiella* genomes.** Data shown summarize Kleborate results for 11,277 non-redundant *Klebsiella* genomes publicly available as at 17 July 2020 (Supplementary Data 2). From left to right: barplots showing source information by geographical region and sample type (coloured as per inset legend); heatmaps showing the prevalence of virulence loci (blue) and predicted AMR drug classes (red) (as per inset scale bars). Genomes are summarized 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  $\beta$ -lactamases, inhR  $\beta$ -lactamase inhibitor, ESBL extended-spectrum  $\beta$ -lactamases, ESBL\_inhR extended-spectrum  $\beta$ -lactamase with resistance to  $\beta$ -lactamase inhibitors, Carb carbapenemase, Bla\_chr intrinsic chromosomal  $\beta$ -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.

KL15, KL2, KL112, KL107, KL151) and accounted for just 34% of the remaining genomes.

**Global population snapshot of *K. pneumoniae* AMR and virulence.** We applied Kleborate to analyse  $n = 13,156$  *Klebsiella* genomes (see “Methods”, Supplementary Data 2). Here we provide a brief overview of the data followed by an exploration of AMR, virulence and the phenomenon of convergence, with the aim to highlight the rich information and types of inferences that can be derived from Kleborate’s output.

The genome data represented isolates collected from a range of sources in 99 countries between 1920–2020 (see Supplementary Data 4, although human isolates from the USA, China and UK dominated the data set accounting for  $n = 4702$  genomes, 35.7% of total). The majority of these genomes were sourced from RefSeq, and among these Kleborate identified 1.0% ( $n = 103/10,747$ ) as a species other than the taxon recorded in NCBI; this is consistent with other studies and highlights the current confusion around taxonomic designations in *Klebsiella*. The most common species was *K. pneumoniae* ( $n = 11,259$ , 86%); the rest comprised other KpSC species (9.4%), other members of the *K. oxytoca* species complex (3.1%) and *K. aerogenes* (1.9%) (Fig. 4, Supplementary Data 4). AMR and virulence genes were concentrated in the KpSC and particularly *K. pneumoniae* (Fig. 4, Table 2).

The collection captured extensive phylogenetic diversity across the *K. pneumoniae* species (see interactive phylogeny at <http://microreact.org/project/bQmTjFQmiCpFBjhoacaL8u>), and Kleborate assigned these genomes to  $\geq 1452$  different STs (1119 known STs across and at least 333 novel STs). Notably, 600 STs (41%) were represented by just a single genome each (accounting for

5.3% of all genomes). We detected  $n = 4$  ST67 (subspecies *rhinoscleromatis*) and  $n = 3$  ST90 (subspecies *ozanae*). A small number of STs were overrepresented, reflecting the bias towards sequencing MDR and hypervirulent isolates, as well as those causing hospital outbreaks. For example, 1354 genomes (12.0%) represented the KPC-associated ST258, which is known to dominate carbapenem-resistant *K. pneumoniae* in the USA and southern Europe (where it has been the subject of intense genomic investigations) but is comparatively rare in other regions of the world<sup>16</sup>. To reduce the impact of these sampling biases in public genome collections, we down-sampled to a non-redundant set of 9705 *K. pneumoniae* genomes representing unique combinations of ST, genetic subcluster (Mash distance  $< 0.0003$ ), virulence genotype, AMR genotype, specimen type, location and year of isolation (see “Methods”). However, we cannot fully correct for the sampling biases inherent in the public genome data and even after subsampling, the 30 most common STs accounted for 63.4% of genomes ( $n \geq 50$  genomes each,  $n = 6151$  total; see Supplementary Fig. 5). Figure 5 shows the distribution of AMR and virulence scores amongst non-redundant genomes from these 30 common *K. pneumoniae* STs ( $n > 50$  per ST), each of which displays high rates of AMR and/or virulence. Importantly, it also highlights the high rates of AMR and/or virulence in some clones (e.g. ST11, ST231), which may correspond to AMR-virulence convergence within a single strain (see Facilitating detection of AMR-virulence convergence section below).

**AMR determinants.** SHV  $\beta$ -lactamases conferring intrinsic resistance to the penicillins were detected in 85.9% of the 9,705 non-redundant *K. pneumoniae* genomes (ESBL forms of SHV were detected in 10.0%). Acquired AMR was widespread (77.1% of genomes had at least one gene or mutation conferring

**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	<i>K. pneumoniae</i>	9705	Ybt: 4309, 44% Clb: 794, 8% Iuc: 1090, 11% Iro: 683, 7% Rmp: 782, 8% RmpA2: 716, 7%	4634, 48%	4173, 43%	
	<i>K. quasipneumoniae</i> subsp. <i>quasipneumoniae</i>	119	-	31, 26%	29, 24%	
	<i>K. quasipneumoniae</i> subsp. <i>similipneumoniae</i>	363	Ybt: 8, 2% Iuc: 6, 2% Iro: 4, 1% Rmp: 4, 1% RmpA2: 3, 0.8%	138, 38%	32, 9%	
	<i>K. quasivariicola</i>	16	-	3, 19%	-	
	<i>K. africana</i>	1	-	-	-	
	<i>K. variicola</i> subsp. <i>variicola</i>	498	Ybt: 15, 3% Iuc: 4, 0.8% Iro: 5, 1% Rmp: 4, 0.8% RmpA2: 2, 0.4%	52, 10%	36, 7%	
	<i>K. variicola</i> subsp. <i>tropica</i>	18	-	2, 11%	1, 6%	
	K. oxytoca species complex	<i>K. oxytoca</i>	98	Ybt: 96, 98%	9 <sup>a</sup> , 9%	6, 6%
		<i>K. grimontii</i>	75	Ybt: 41, 55%	1 <sup>a</sup> , 1%	3, 4%
		<i>K. huaxiensis</i>	4	-	1 <sup>a</sup> , 25%	-
<i>K. michiganensis</i>		144	Ybt: 102, 71% Clb: 1, 0.7%	21 <sup>a</sup> , 15%	33, 23%	
NA	<i>K. pasteurii</i>	21	Ybt: 21, 100%	1 <sup>a</sup> , 5%	-	
	<i>K. spallanzanii</i>	4	-	- <sup>a</sup>	-	
	<i>K. aerogenes</i>	209	Ybt: 101, 48% Clb: 95, 45% Iro: 190, 91%	24, 11%	34, 16%	
	<i>K. indica</i>	2	-	-	-	

Ybt yersiniabactin, Clb colibactin, Iuc aerobactin, Iro salmochelin, Rmp hypermucoidy conferred by rmpADC locus.

<sup>a</sup>Excluding OXY genes that are conserved in *K. oxytoca* species complex.

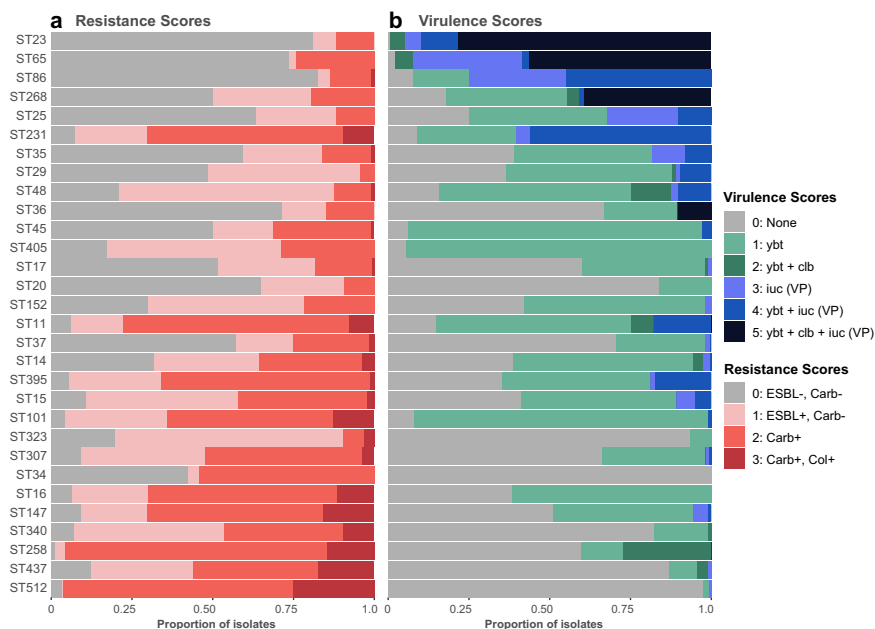
acquired AMR detected) and 71.6% of genomes were predicted to be MDR (acquired resistance to  $\geq 3$  drug classes<sup>50</sup>), a much higher rate than is reported in most geographical regions<sup>3,51–53</sup>, reflecting the bias within public genome collections. The majority of genomes had a non-zero resistance score, reflecting the presence of ESBL and/or carbapenemase genes: 22.3%, 37.1% and 5.9% genomes had resistance scores of 1, 2 and 3 respectively. Mean resistance scores increased through time (barplot in Fig. 1b), and were correlated with the annual prevalence of ESBLs (Spearman correlation coefficient  $r_s = 0.80$ ) and carbapenemases ( $r_s = 0.98$ ) (line plots in Fig. 1b). This trend could be an artefact of sampling bias towards the selective sequencing of AMR isolates, however it is consistent with the increasing AMR rates reported in surveillance studies globally<sup>54–56</sup>.

Comparatively higher prevalence of acquired AMR genes was observed in some STs (Supplementary Fig. 5). Many of these STs represent recognized MDR clones largely from clinical samples that were also associated with high mean resistance scores (Fig. 6a, b), driven by high frequency of ESBL and carbapenemase genes (Fig. 5, Supplementary Fig. 6a, b). The most common ESBLs/carbapenemases were widely detected across the population (46–299 STs each), including amongst the top 30 common STs (prevalence range per ST, 0.1–100%; see Supplementary Fig. 6a, b), highlighting their mobile nature. The notable exception was CTX-M-65, which appeared to be largely clone-specific, detected in only 9 STs and ST11 accounting for 96.7% of these genomes.

Colistin resistance determinants<sup>57–59</sup> were detected in 8.7% of the non-redundant *K. pneumoniae* genomes. These were mostly nonsense mutations in MgrB or PmrB (83.5%) rather than acquisition of an *mcr* gene (15.8%), and an additional 6 genomes with both acquired *mcr* and truncated MgrB/PmrB). The rate of detection ranged from 0 to 25.2% for the 30 most common STs, and was highest amongst ST512, ST437, ST147, ST16 and ST258 (Supplementary Fig. 6c), each of which are also associated with high rates of carbapenem-resistance. Porin mutations were detected in 37.9% of genomes (34.0% OmpK35, 20.2% OmpK36, 16.3% both). High prevalence of specific porin defects have been reported previously in some clones<sup>41,42</sup>, and this was reflected in our analysis of ST258 and its derivative ST512. We observed OmpK35 truncations in 99.9% of non-redundant ST258 genomes (with or without truncations or substitutions in OmpK36), and truncations in OmpK35 and/or OmpK36 in all ST512 (99.4% with OmpK35 truncations, 94.4% with the OmpK36GD mutation, see Supplementary Fig. 6d).

**Virulence loci.** The prevalence of acquired siderophores and colibactin loci amongst non-redundant *K. pneumoniae* genomes was 44.4% *ybt*, 7.5% *clb*, 11.2% *iuc* and 7.0% *iro*. The loci were found across diverse *K. pneumoniae* STs (391 STs with *ybt*, 56 with *clb*, 144 with *iuc*, 108 with *iro*) but were rarely detected in other *Klebsiella* species (with the exception of *ybt* among the *K. oxytoca* species complex, see Fig. 4) indicating frequent mobilisation within *K. pneumoniae* but not between species (Supplementary Data 5,





**Fig. 5** Distribution of (a) resistance and (b) virulence scores among genomes belonging to the 30 most common *K. pneumoniae* lineages. Data shown summarize Kleborate results for non-redundant set of 9705 publicly available *K. pneumoniae* genomes (Supplementary Data 2). 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  $\beta$ -lactamase, Carb carbapenemase, Col colistin resistance determinant.

Supplementary Fig. 7). Mean virulence scores increased through time (barplot in Fig. 1a) and were correlated with an annual prevalence of *ybt* ( $r_s = 0.88$ ), *iuc* ( $r_s = 0.76$ ) and *rmp* ( $r_s = 0.62$ ) (line plots in Fig. 1a). Figure 5b shows the frequency of virulence scores in the top 30 most common STs in the non-redundant genome set. Sixteen of these common STs had  $\geq 40\%$  of genomes carrying the ICEKp-associated *ybt* without the virulence plasmid-associated *iuc* locus (i.e. virulence score = 1–2), including well-known MDR clones ST258, ST11, ST14, ST15, ST101, ST147, ST152, ST395. Only the hvKp clones (ST23, ST86, ST65) and ST231 had a high frequency of *iuc* (virulence score  $\geq 3$ ).

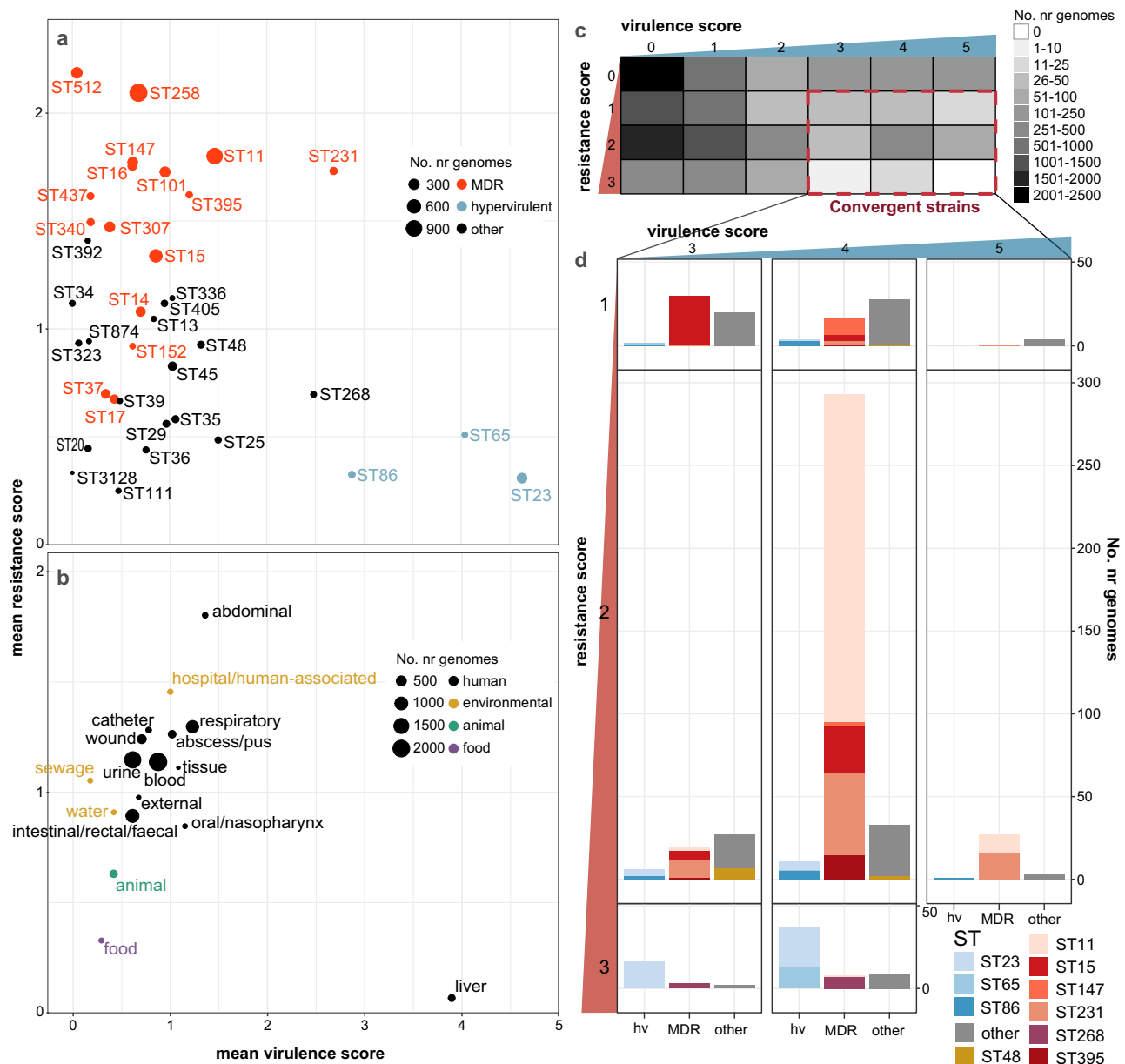
In addition to detecting the presence of virulence loci, Kleborate reports on their completeness, genetic lineages and associated MGE variants, which can provide insights into their dissemination. Most of the virulence loci identified in the non-redundant *K. pneumoniae* data set (98%) matched one of the genetic lineages described previously<sup>34,35</sup> (Supplementary Data 5). Supplementary Fig. 8a shows the frequency of *iuc* lineages in *K. pneumoniae* STs with  $\geq 20$  non-redundant genomes and at least one genome harbouring *iuc*. There were four STs for which  $>60\%$  genomes harboured *iuc*, and only a single *iuc* lineage was detected in each (*iuc1* in ST23, ST65, ST86; *iuc2A* in ST82), consistent with the long-term persistence of a specific virulence plasmid in these well-known hypervirulent clones. In contrast, *iuc* was less frequent among other STs, several of which were associated with multiple *iuc* lineages (e.g. ST231, ST25, ST35), consistent with more recent and/or transient virulence plasmid acquisitions (mostly *iuc1*, followed by *iuc3* and *iuc5*).

Frameshift mutations (i.e. truncations) and/or incomplete loci (i.e. missing at least one gene) were detected in 10%, 28.5%, 13.6% and 17.7% of non-redundant *K. pneumoniae* genomes with *ybt*, *clb*, *iuc* and *iro* respectively (Supplementary Data 6). While some of these may erroneously arise from contig breaks in draft genome assemblies, true truncations or missing genes may reflect a lack of function. The latter is likely true for instances where we observe conserved frameshift mutations across entire lineages, e.g.

frameshift mutations were detected in *iucA* for all *iuc3+* genomes and in *iroC* for all *iro3+* and *iro4+* genomes.

The hypermucoidity locus *rmpADC* was detected in 8.4% of non-redundant *K. pneumoniae* genomes (and just eight genomes of other KpSC species, Supplementary Data 5). The majority of these genomes (67.2%, belonging to  $>79$  STs) carried intact copies of all three genes, thus likely express the hypermucoidity phenotype. Intact *rmpADC* was common in *iuc*-positive genomes of the hvKp clones ST23 and ST86, as well as MDR clones ST29 and ST101 (Supplementary Fig. 8b). Many other *iuc*-positive genomes carried *rmpADC* loci with truncated or missing genes, which likely do not confer the hypermucoidity phenotype. Notably, these included hvKp clones ST65 and ST82, as well as MDR clones ST231, ST15 and ST14. The *rmpA2* gene was detected in 7.4% of non-redundant *K. pneumoniae* genomes, but was mostly present in truncated form (89.0% of *rmpA2+* genomes) due to frameshifts within a poly-G tract<sup>60</sup>. The latter highlights the importance of considering not only the presence/absence of a given gene, but also whether it encodes a full-length protein, which may have important clinical implications.

**Facilitating detection of AMR-virulence convergence.** AMR and virulence determinants have until recently been segregated in non-overlapping *K. pneumoniae* populations<sup>14,19</sup>, as clearly indicated by the distributions of AMR and virulence scores among STs (Figs. 5, 6a). However, reports of convergent AMR-virulent strains with the potential to cause difficult-to-treat infections are increasingly common<sup>16,61</sup>. Kleborate facilitates rapid identification of such strains on the basis of resistance and virulence scores (convergence defined as virulence score  $\geq 3$  and resistance score  $\geq 1$ , Fig. 6c). Based on these scores, we observed a total of 601 convergent *K. pneumoniae* (510 non-redundant) with the highest proportion corresponding to a virulence score of 4 (indicative of yersiniabactin plus aerobactin/virulence plasmid detection) and resistance score of 2 (carbapenem resistance).

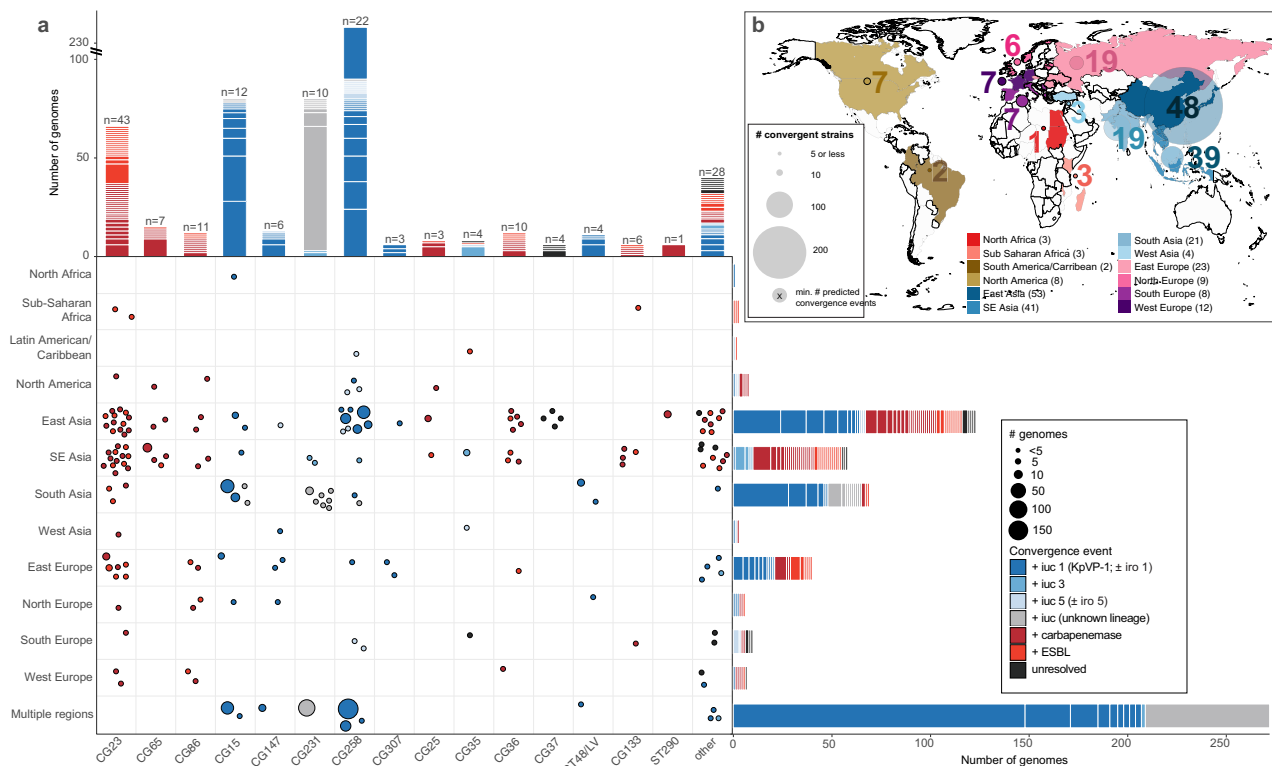


**Fig. 6 Insights from resistance and virulence scores.** Data shown summarize Kleborate results for non-redundant set of 9,705 publicly available *K. pneumoniae* genomes (Supplementary Data 2). **a–b** Mean resistance and virulence scores grouped by **(a)** lineage and **(b)** sample type. Each circle 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  $\geq 3$  (carrying *iuc*) and resistance score of  $\geq 1$  (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 others; and coloured by ST (as per inset legend).

The majority of convergent genomes (74.5%) were concentrated within a small number of STs comprising well-known hypervirulent (e.g. ST23, ST86, ST65) or MDR lineages (e.g. ST11, ST15, ST231 and ST147) (Figs. 6c, d, 7). Using the combination of genotyping data and a Mash-distance-based neighbour-joining tree (<http://microreact.org/project/JDyan46yctyDh6weEUjWN>) we identified 174 unique convergence events (Supplementary Data 7, see “Methods” for details). The majority of convergence events ( $n = 95$ , 55%; 137 genomes) were attributed to the acquisition of AMR in strains already carrying *iuc* (including 65 events (37.4%) corresponding to the acquisition of AMR by known hypervirulent clones; see Fig. 7, Supplementary Data 7). A further 67 events (38.5%; 449 genomes) were attributed to the acquisition of *iuc* in strains already

carrying ESBLs/carbapenemases, and for 12 events (15 genomes) the order of acquisition could not be resolved.

The most common virulence plasmid, KpVP-1 (*iuc1*  $\pm$  *iro1*), accounted for 63% of virulence plasmid acquisition events ( $n = 42$  acquisitions), while *iuc3* plasmids, the *E. coli* derived *iuc5* ( $\pm$  *iro5*) and *iuc/iro* unknown (i.e. novel or divergent *iuc/iro* loci) accounted for 8%, 13% and 16%, respectively (Fig. 7). AMR acquisitions by hypervirulent clones involved the ESBL/carbapenemase genes that are most common in the general *K. pneumoniae* population: KPC-2 (24%), OXA-232 (17%) and CTX-M-15 (16%). The majority of convergence events (87%) were associated with just a small number of genomes (i.e.  $n \leq 3$ ); however, five events were associated with  $>20$  genomes in the



**Fig. 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 Supplementary Data 7, assignment of genomes to events in Supplementary Data 2). Circles are scaled by the number of total genomes linked to the event and coloured to indicate whether convergence is inferred to have occurred via (i) acquisition of AMR gene/s (ESBL or carbapenemase/s) by a hypervirulent lineage, (ii) acquisition of an *iuc*-encoding plasmid by an AMR or non-AMR lineage, or (iii) unresolved means as per inset legend. Marginal barplots show the number of convergence events (colour blocks) and genomes (black 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 coloured on the map; circles represent the number of convergent genomes detected in each UN-defined geographical region (indicated by colour, 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.

complete dataset, which may indicate clonal expansion and dissemination of the corresponding convergent strains locally and/or between countries. One such event corresponded to the ST11-KPC + KpVP-1 deletion variant strain that was originally reported in 2017<sup>20</sup> and has since been recognized as widely distributed in China<sup>20–24</sup>. The complete public genome set (i.e. counting redundant genomes) included 148 genomes corresponding to this specific ST11 convergence event mostly from China but also from France ( $n = 2$ ). Notably though, this was only one of 50 convergence events that we detected in China, including 8 involving the acquisition of *iuc1* or *iuc5* by ST11 (see Supplementary Data 7, and interactive tree at <http://microreact.org/project/JDyan46yctyDh6weEUjWN>). Additional events associated with >20 genomes included (i) ST231-MDR + virulence plasmids carrying novel *iuc* lineages detected in India, Pakistan, Switzerland, Thailand and USA, (ii) ST15-CTX-M-15 + KpVP-1 in Pakistan, (iii) ST15-MDR + KpVP-1 in China and Nepal, and (iv) another distinct ST11-KPC-2 + KpVP-1 event in China. Including the above three examples, 11 convergence events appeared to involve intercountry expansion of which one has been previously documented<sup>62</sup>.

Overall, convergent genomes were detected originating from most geographical regions for which genome data was available,

but some regions had many more events than others (Fig. 7, Supplementary Data 7). This uneven distribution may stem from a skew in the number of genomes available per region (e.g. due to variation in accessibility or application of genome sequencing). Nevertheless, the number of convergent genomes in the eastern, southeastern and southern parts of Asia were noticeably high, driven by the frequency of convergence events detected in China ( $n = 50$  events) and Thailand ( $n = 26$  events) as well as putative clonal expansions of these strains as discussed above (Fig. 7). Of note, AMR acquisitions by hypervirulent lineages were particularly frequent within East and Southeast Asia where hypervirulent infections are most frequently reported, alongside countries from eastern and northern Europe.

Outside of *K. pneumoniae*, convergence events were rare: we detected  $n = 2$  *K. quasipneumoniae* subsp. *similipneumoniae* (ST367 with KpVP-1 and CTX-M-15; ST3387 with *iuc3* and CTX-M-55) and  $n = 2$  *K. variicola* subsp. *variicola* (ST595 with KpVP-1 and KPC-2; ST1848 with *iuc5* and KPC-2).

**Genotyping *K. pneumoniae* from metagenome data.** There is increasing interest in detection and typing of *K. pneumoniae* direct from gut metagenome data<sup>63</sup>, due to the role of *K.*

*pneumoniae* gut colonization as a source of acute infections and as a contributor to chronic diseases<sup>7,8</sup>. We tested Kleborate's performance by application to  $n = 40$  metagenomes from which at least one KpSC isolate was cultured and sequenced, as part of the Baby Biome Study<sup>64</sup>. We compared the results of running Kleborate on metagenome-assembled genomes (MAGs, i.e., species-specific contig bins extracted from whole-metagenome assemblies) vs. KpSC isolate whole-genome sequence(s) cultured from the same fecal sample. Thirty-two metagenomes had >1% relative abundance of KpSC, and genotyping of MAGs from these yielded results consistent with genotyping of cultured isolates for 26/32 samples (16 with identical genotypes reported for species, ST, K/O locus, virulence and AMR; 10 with close matches; see Supplementary Fig. 8, Supplementary Data 8–9). As expected, MAG-derived genotypes were closest to those of isolates when only one KpSC strain was cultured from the sample (see Supplementary Fig. 8, Supplementary Data 9). We were unable to assess the reliability of Kleborate for distinguishing KpSC from non-KpSC *Klebsiella* in a single metagenome sample, as there were no samples with sufficient abundance of both KpSC and non-KpSC species to yield the corresponding MAGs for both (i.e. in the  $n = 9$  samples for which both KpSC non-KpSC were isolated, the relative abundance of one species or the other was too low to yield a corresponding MAG). Kleborate analysis of whole metagenome assemblies (as opposed to individual MAGs) is not recommended: species detection and ST assignment matched that of the corresponding WGS isolates for only  $n = 4/40$  metagenome assemblies, which is unsurprising as the whole metagenomes include sequences derived from dozens of different bacteria, many of which harbour homologs of genotyping targets.

## Discussion

WGS is being increasingly implemented in research and public health labs as a cost- and time-effective option for tracking pathogens and AMR determinants. However, the identification of known clinically relevant features remains a key bottleneck that hinders widespread adoption of genome surveillance. We have presented a comprehensive framework and tool for rapid genotyping of *Klebsiella* species genomes: Kleborate is a single unified approach for species detection, MLST and genotyping of key virulence and AMR determinants. It focuses only on genomic features for which there is strong evidence of a clinically relevant phenotype in KpSC and presents the data in a readily interpretable format, with numerical summaries and categorical scores corresponding to measures of potential clinical risk. While there is generally high concordance between genotypes and phenotypes (e.g. AMR gene detection is typically predictive of phenotype<sup>65–69</sup>), Kleborate reports only genotypes and does not provide predictions of clinical antibiotic resistance or virulence. It is also important to note that in some instances, accurate phenotypic predictions may not be possible given the complexities in the underlying genetic mechanisms that are not yet fully understood (e.g. the interactions of specific carbapenemases and OmpK variants and how these contribute to MIC for carbapenems).

A key strength of the Kleborate framework is its species-specific approach. This is particularly important for accurate interpretation of AMR and virulence gene screens from WGS, wherein the use of generic databases and tools can result in confusion. Notable examples include the intrinsic *oqxAB* and *fosA* alleles, which unlike for other Enterobacterales, do not confer resistance to quinolones and fosfomycin when expressed in KpSC. Kleborate does not report these intrinsic alleles, neither does it report intrinsic virulence determinants such as the siderophore enterobactin, which is known to play a role in KpSC

pathogenicity but for which the presence alone cannot be considered to indicate enhanced virulence of one isolate over another. Correct taxonomic identification of *K. pneumoniae* can be difficult in itself, hence the inbuilt speciation tool is an important feature (and here identified nearly 100 RefSeq genomes with incorrect species/subspecies assignments).

Another strength of our approach is the rich data output by Kleborate, which facilitates in-depth investigation of population structure, AMR and virulence epidemiology. This allows rapid exploration and understanding of (i) hypervirulence-associated loci and the molecular drivers of their dissemination (Supplementary Figs. 5 and 8); (ii) molecular mechanisms of complex AMR phenotypes e.g. carbapenem resistance (Fig. 3); (iii) AMR and virulence trends (Figs. 1, 5 and 6); (iv) emerging convergent AMR-virulent strains so that they can be targeted for surveillance and infection control (Fig. 7); (v) overrepresented STs and genotypes, which may be indicative of transmission clusters that should be targeted for further investigation (as demonstrated for the EuSCAPE surveillance genomes, Fig. 2a); (vi) surface antigen epidemiology, which can inform the design of novel vaccines and therapeutics (Fig. 2b, c). Notably Kleborate can also yield useful genotyping results from metagenomics data (Supplementary Fig. 9), which is gradually being adopted for clinical and surveillance applications relevant to *K. pneumoniae*. User interpretation of Kleborate's extensive data output can be guided by the accompanying web-based visualization app, Kleborate-Viz. Through this app, many of the analyses and plots presented in this manuscript can be rapidly replicated, and further explored in an interactive manner.

Kleborate is designed to facilitate detection and tracking of clinically relevant AMR and virulence determinants from genome data, and analysis of public data not only identified specific clones and genes associated with one or the other of AMR and virulence (Figs. 5, 6), but also 601 genomes in which the two converge (carrying *iuc+* virulence plasmids and ESBL and/or carbapenemase genes; Fig. 7). We estimated at least 174 unique AMR-hypervirulence convergence events; the majority were detected within a single isolate ( $n = 121$  events), but many others appear to be associated with local outbreaks or larger-scale spread and apparently across multiple countries (Supplementary Data 7). Some of the convergence events in China and other countries in the neighbouring South and Southeast Asia regions have been extensively reported<sup>16,20,51,61</sup>, but to our knowledge, a significant number had not been recognized previously. These include ST231-MDR (most with OXA-232, remainder with ESBL only) + *iuc*, which has been reportedly circulating in India<sup>51</sup>, and our analysis also detected in Pakistan, Thailand, Switzerland and USA.

Kleborate has already been widely adopted by the *Klebsiella* research community—at least 74 studies have reported using the Kleborate software package, including larger-scale genome surveillance studies in South and Southeast Asia, the Caribbean and the United States<sup>31,51,52</sup> (full list in Supplementary Data 10). Kleborate is freely available as a standalone command-line tool for local high-throughput analyses or incorporation into existing bioinformatics workflows (<https://github.com/katholt/Kleborate>), and can be easily accessed through the online tool Pathogenwatch (<https://pathogen.watch/>). With such broad accessibility and utility, Kleborate is poised to become a cornerstone of the *Klebsiella* genomic surveillance toolkit that can help inform containment and control strategies targeting this priority pathogen. Input will be sought from the *Klebsiella* and AMR surveillance communities to guide ongoing development, including establishing formal criteria for inclusion of AMR or virulence features and improving reporting of results.

## Methods

**Kleborate software: implementation and genotyping logic.** Kleborate (v.2.0.0) is a command-line tool written in Python and is freely available under the GNU v3.0 license at <http://github.com/katholt/Kleborate>. It takes as input one or more whole-genome assemblies (FASTA format), types each one against a series of screening databases outlined in detail below, and returns results in a tab-delimited text file (one genome per row). On default settings, Kleborate will report assembly quality metrics, taxonomic assignment, MLST and virulence loci genotypes. Screening for AMR determinants, and/or K/O serotyping via Kaptive<sup>36</sup>, is optional (Table 1).

**Assembly quality.** Assembly quality metrics, reported to help users assess the reliability of genotyping results, are: contig count, contig N50, largest contig size, total genome size, and number of ambiguous bases (e.g. 'N'). Low-quality warnings are flagged if (i) ambiguous bases are detected; (ii) assembly length falls outside the expected range of 4.5–7.5 Mbp; or (iii) N50 is below 10,000 bp. Users should carefully consider the genotyping outputs for low-quality assemblies.

**Taxonomic assignment.** Kleborate's species prediction function provides a convenient way to confirm species, including differentiating between the closely related members of the KpSC which are frequently misclassified using laboratory techniques. Kleborate calculates Mash<sup>40</sup> distances between the input genome/s and a curated collection of reference assemblies from different *Klebsiella* and other Enterobacterales, and reports the species with the smallest distance. Mash distance  $\leq 0.02$  is reported as a strong match,  $\leq 0.04$  as weak (only when no strong matches are found, see Supplementary Note 1 for further details).

**MLST.** Genomes assigned to species in the KpSC are assigned STs using nucleotide BLAST against the established *K. pneumoniae* chromosomal seven-locus MLST scheme<sup>29</sup> described and maintained on the *K. pneumoniae* BIGSdb site hosted at the Pasteur Institute (<http://bigsdbs.pasteur.fr/klebsiella/klebsiella.html>).

**Virulence gene detection and typing.** Virulence loci (*ybt*, *iuc*, *iro*, *clb*, *rmpADC*, *rmpA2*) are detected using nucleotide BLAST search against the database of known alleles. The best hit allele for each gene (with  $\geq 90\%$  identity and  $\geq 80\%$  coverage) is reported in the main virulence columns. If the majority of genes expected for the locus are present, then the alleles are used to calculate STs which are reported along with their associated lineage and MGE (based on previously defined schemes: YbST for *ybt*, CbST for *clb*, AbST for *iuc*, SmST for *iro*, according to the previously defined schemes<sup>34,35</sup>; and a novel RmST scheme for the *rmpADC* locus). To generate the RmST typing scheme we used the same 2733 genomes from our original virulence plasmid study<sup>35</sup> to screen and extract the sequences for *rmpADC* and define allele numbers and STs. These ST sequences cluster into four distinct lineages associated with distinct MGEs (*rmp1* with KpVP-1, *rmp2* with KpVP-2, *rmp2A* with the *iuc2A* virulence plasmids, and *rmp3* with ICEKp1; to be described in detail elsewhere). Where the best hit for a gene is a weak match (80–90% identity, 40–80% coverage) this is reported in the 'spurious hits' column. Truncations are detected by translating the best-matching nucleotide sequence for each query gene into amino acids and comparing to the reference length (expressed as % amino acid length from the start codon, those  $< 90\%$  are reported).

The presence of *ybt*, *clb* and *iuc* are used to assign a virulence score as follows: 0 = none present, 1 = yersiniabactin only, 2 = colibactin without aerobactin (regardless of yersiniabactin, however, *ybt* is almost always present when *clb* is), 3 = aerobactin only, 4 = aerobactin and yersiniabactin without colibactin, and 5 = all three present. The presence of *iro* (salmochelin) is not used to calculate the virulence score because its presence is very strongly associated with aerobactin. The scoring aims to capture the general hierarchy of virulence and associated loci that have emerged from the literature over the last two decades. Yersiniabactin facilitates immune escape (by evading Lcn2) and has been shown to increase virulence in multiple strain backgrounds<sup>70</sup>, however it is not associated with toxinogenic activity (like *clb*<sup>71</sup>) or growth in blood (like *iuc*<sup>72</sup>). Ybt is also the most common virulence determinant (~20–50% of clinical isolates) and is almost always present in *clb*+ isolates (*clb* is carried by one of the 14 forms of the yersiniabactin ICE described to date<sup>34</sup>) and in isolates carrying the virulence plasmid<sup>16,35</sup>; hence the presence of yersiniabactin is assigned a score of 1. The next increment in score (2) is assigned to the presence of the genotoxin *clb* in addition to *ybt*, because of *clb*'s genotoxic activity against mammalian cells<sup>71</sup> but there is only limited evidence that it elevates virulence substantially in the absence of the virulence plasmid. A recent study evaluating markers of hypervirulence identified the *K. pneumoniae* virulence plasmid markers (*iuc*, *iro*, *rmpA*, *rmpA2*, *peg-344*) as being highly diagnostic of hypervirulent vs classical *K. pneumoniae* infection amongst human clinical isolates, and are also predictive of mortality in a murine sepsis model<sup>73</sup>. These virulence plasmid markers are in very strong genetic linkage (~99%), hence in principle any of these would serve as a good marker for the next increment in virulence score; however, we selected *iuc* as it is the most clearly and directly functionally related to sepsis (promoting growth in blood via the acquisition of iron from transferrin<sup>72,74</sup>). While there is limited data with which to assess the individual contributions of *ybt*, *clb* and the virulence plasmid when present in combination, it is logical to score *iuc* + *ybt* higher than *iuc* without *ybt*. In the well-described hypervirulent clone ST23, the most widely distributed sublineage that is responsible for the majority of liver abscess documented globally (CG23-I) is

distinguished by carrying *clb*, compared to the rest of the ST23 population which carries *clb*-ICEs and are rarely seen<sup>75</sup>; hence we assign the presence of *iuc* (virulence plasmid) + *ybt* + *clb* as the highest score.

**Detection and typing of antimicrobial resistance determinants.** When AMR detection is switched on, Kleborate screens for known acquired AMR determinants using a curated version of the CARD AMR nucleotide database (v3.0.8 downloaded February 2020; see [doi.org/10.6084/m9.figshare.13256759.v1](https://doi.org/10.6084/m9.figshare.13256759.v1) for full details on curation). Genes are identified using nucleotide BLAST (and amino acid search with tBLASTx if no exact nucleotide match is found). Gene truncations and spurious hits are identified as described above for virulence genes. Unlike the acquired forms, the intrinsic variants of *oqxAB*, chromosomal *fosA* and *ampH* are not associated with clinical resistance in KpSC and are therefore not reported. However, SHV, LEN or OKP  $\beta$ -lactamase alleles intrinsic to KpSC species are known to confer clinical resistance to penicillins and are reported in the *Bla\_chr* column. Acquired SHV variants, and individual SHV sequence mutations known to confer resistance to extended-spectrum  $\beta$ -lactams or  $\beta$ -lactamase inhibitors, are reported separately (see Supplementary Note 3, Supplementary Data 11 and 12 for details).

Chromosomally encoded mutations and gene loss or truncations known to be associated with AMR are reported for genomes identified as KpSC species. These include fluoroquinolone resistance mutations in GyrA (codons 83 and 87) and ParC (codons 80 and 84)<sup>76</sup>, and colistin resistance from truncation or loss of MgrB and PmrB<sup>57–59</sup> (defined as  $< 90\%$  amino acid sequence coverage). Mutations in the OmpK35 and OmpK36 osmoporphins reportedly associated with reduced susceptibility to  $\beta$ -lactamases<sup>41,42</sup> are also screened and reported for KpSC genomes, and include truncation or loss of these genes and OmpK36GD and OmpK36TD transmembrane  $\beta$ -strand loop insertions<sup>41</sup>. SHV  $\beta$ -lactamase, GyrA, ParC and OmpK mutations are identified by alignment of the translated amino acid sequences against a reference using BioPython, followed by an interrogation of the alignment positions of interest (see Supplementary Note 3, Supplementary Data 11 and 12 for a list of relevant positions).

AMR genes and mutations are reported by drug class, with  $\beta$ -lactamases further categorized by enzyme activity ( $\beta$ -lactamase, ESBL or carbapenemase, with/without resistance to  $\beta$ -lactamase inhibitors). Horizontally acquired AMR genes are reported separately from mutational resistance and contribute to the AMR gene count; these plus chromosomal mutations count towards the number of acquired resistance classes (intrinsic SHV alleles, reported in *Bla\_chr* column, are not included in either count). Resistance scores are calculated as follows: 0 = no ESBL or carbapenemase, 1 = ESBL without carbapenemase (regardless of colistin resistance); 2 = carbapenemase without colistin resistance (regardless of ESBL); 3 = carbapenemase with colistin resistance (regardless of ESBL).

**Serotype prediction.** By default, genomes are screened against the *wzi* database in the *Klebsiella* BIGSdb (using nucleotide BLAST) which is used to predict capsule (K) type based on a previously defined scheme<sup>77</sup>. This allows rapid typing however the relationship between *wzi* allele and K type is not one-to-one<sup>36</sup>. If surface antigen prediction is important to users they can obtain more robust identification of K and O antigen (LPS) loci by switching on serotype prediction with Kaptive<sup>36</sup> (—kaptive), which adds a few minutes per genome to Kleborate's runtime.

**Data visualization.** To facilitate interpretation of Kleborate's rich data output we provide a web-based application (Kleborate-Viz, <https://kleborate.erc.monash.edu/>), implemented in R Shiny, which takes as inputs a Kleborate results file (required), sample metadata (CSV format, optional) and MIC data (CSV format, optional). User data is temporarily stored on the server for the duration of the session and is immediately deleted when the session is terminated upon closing the browser window or tab.

**Genome analysis.** The analyses reported here result from applying Kleborate v2.0.0 ([doi:10.5281/zenodo.4923015](https://doi.org/10.5281/zenodo.4923015)) to publicly available genome collections. A total of 13,156 *Klebsiella* WGS assemblies, encompassing non-duplicate isolates with unique BioSample accessions identified from published studies (some deposited as read sets only, which were assembled using Unicycler v0.4.7<sup>78</sup>, data sources summarized in Supplementary Data 13) plus any additional genomes designated as *Klebsiella* in NCBF's RefSeq repository of genome assemblies (as of 17 July 2020). In order to minimize the impact of sampling bias favoring common MDR and/or virulent lineages and those causing outbreaks, we subsampled the collection into a 'non-redundant' dataset of 11,277 genomes (9705 *K. pneumoniae*) as follows. Pairwise Mash distances were calculated using Mash v2.1, and used to cluster genomes using single-linkage clustering with a threshold of 0.0003. These clusters were further divided into non-redundant groups with unique combinations of (i) Mash cluster, (ii) chromosomal ST, (iii) virulence gene profiles (i.e. presence of *ybt/clb/iro/iuc* loci and lineage assignment), (iv) AMR profiles, (v) year and country of isolation, and (vi) specimen type where available. For each resulting non-redundant group, one genome was selected at random as the representative for analyses. The full list of genomes, including database accessions, isolate information, cluster/group assignment, and Kleborate results are provided in Supplementary Data 2. The subset of 1624 *K. pneumoniae* assemblies deposited in RefSeq by the European ESCAPE surveillance study<sup>33</sup> (out of 1649 reported in original study; Supplementary Data 2) were used for the ESCAPE analyses reported in

Figs. 2 and 3. The Kleborate-Viz web application is pre-loaded with the non-redundant and EuSCAPE WGS datasets reported in this paper, and can be used to reproduce the plots shown in Figs. 1a–c, 2b, c, 3, 6a, b and to further explore the Kleborate results.

**Analysis of AMR-virulent convergent genomes.** The Mash-distance-based neighbour-joining tree (<https://microreact.org/project/JDyan46cYtDh6weEUjWN>) was used to identify unique subtrees of AMR-virulent convergent genomes (i.e. carrying *iuc* plus an ESBL and/or carbapenemase gene), which revealed 174 unique clusters representing independent convergence events. The majority of these were within known hypervirulent ( $n = 65$ ) or MDR clones ( $n = 57$ ; see Supplementary Data 7); for these the order of acquisition was trivially assigned as virulence then resistance, or resistance then virulence, respectively. For the remaining  $n = 52$  events, the distribution of resistance and *iuc* genes in each subtree and its sister clades was manually inspected by two independent analysts to infer the order of acquisition based on the maximum parsimony principle. In 12 cases, the order of virulence vs. AMR acquisition could not be resolved (see Supplementary Data 7).

**Metagenome analysis.** We downloaded metagenomic reads, and matched isolate WGS assemblies, for  $n = 47$  infant gut microbiota samples deposited by the Baby Biome Study<sup>64</sup>. Metagenome reads were assembled using SPAdes version 3.13.1<sup>79</sup> with the `--meta` flag and the resulting contigs binned using MaxBin v2.2.7<sup>80</sup>. Seven metagenomes failed to assemble due to memory and compute walltime constraints, hence we report results for 40 samples (Supplementary Data 9). Kleborate was run separately on the full metagenome assemblies, all contig bins (from which the *Klebsiella* bin could then be identified), and the matched WGS assemblies. Metagenomic read sets were also analysed using Kraken 2.0.7<sup>81</sup> and Bracken v2.5<sup>82</sup> (with a custom GDTB release 89 database<sup>83</sup>) to estimate the relative abundance of KpSC reads in each metagenome.

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

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Supplementary Data 2 lists accession numbers for each genome analyzed in this study, alongside the isolate collection metadata where available and Kleborate-generated output. An interactive phylogeny of all public *Klebsiella* genomes alongside (i) the corresponding Kleborate data or (ii) information relating to convergence events can be accessed at (i) <https://microreact.org/project/bQmTJfQmCpFBjhoacL8u> and (ii) <https://microreact.org/project/JDyan46cYtDh6weEUjWN> respectively. Supplementary Data 8 lists accession numbers for metagenome reads, matched isolate whole-genome assemblies and the Kleborate-generated output.

## Code availability

The code and detailed instruction manuals for running Kleborate and Kleborate-Viz can be found at <http://github.com/katholt/Kleborate> and <https://kleborate.erc.monash.edu/> respectively. The code for Kleborate v2.0.0 used in the study can be found at doi:10.5281/zenodo.4923015<sup>84</sup>.

Received: 4 February 2021; Accepted: 16 June 2021;

Published online: 07 July 2021

## References

- World Health Organisation. *Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics* (2017).
- Wyres, K. L. & Holt, K. E. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Curr. Opin. Microbiol.* **45**, 131–139 (2018).
- Gorrie, C. L. et al. Gastrointestinal carriage is a major reservoir of *K. pneumoniae* infection in intensive care patients. *Clin. Infect. Dis.* **65**, 208–215 (2017).
- Martin, R. M. et al. Molecular epidemiology of colonizing and infecting isolates of *Klebsiella pneumoniae*. *mSphere* **1**, e00261-16 (2016).
- Chung, D. R. et al. Fecal carriage of serotype K1 *Klebsiella pneumoniae* ST23 strains closely related to liver abscess isolates in Koreans living in Korea. *Eur. J. Clin. Microbiol. Infect. Dis.* **31**, 481–486 (2012).
- Lin, Y.-T. et al. Seroepidemiology of *Klebsiella pneumoniae* colonizing the intestinal tract of healthy Chinese and overseas Chinese adults in Asian countries. *BMC Microbiol.* **12**, 13 (2012).
- Kaur, C. P., Vadivelu, J. & Chandramathi, S. Impact of *Klebsiella pneumoniae* in lower gastrointestinal tract diseases. *J. Dig. Dis.* **19**, 262–271 (2018).
- Podschun, R. & Ullmann, U. *Klebsiella* spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clin. Microbiol. Rev.* **11**, 589–603 (1998).
- Petrosillo, N., Taglietti, F. & Granata, G. Treatment options for colistin resistant *Klebsiella pneumoniae*: present and future. *J. Clin. Med.* **8**, 934 (2019).
- Tooke, C. L. et al.  $\beta$ -Lactamases and  $\beta$ -lactamase inhibitors in the 21st century. *J. Mol. Biol.* **431**, 3472–3500 (2019).
- Geneva: World Health Organization. *Prioritization of Pathogens To Guide Discovery, Research and Development of New Antibiotics for Drug-Resistant Bacterial Infections, Including Tuberculosis* (2017).
- Shon, A. S., Bajwa, R. P. S. & Russo, T. A. Hypervirulent (hypermucoviscous) *Klebsiella pneumoniae*: a new and dangerous breed. *Virulence* **4**, 107–118 (2013).
- Siu, L. K., Yeh, K., Lin, J., Fung, C. & Chang, F. *Klebsiella pneumoniae* liver abscess: a new invasive syndrome. *Lancet Infect. Dis.* **12**, 881–887 (2012).
- Wyres, K. L. et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet.* **15**, e1008114 (2019).
- Brisse, S. et al. Virulent clones of *Klebsiella pneumoniae*: Identification and evolutionary scenario based on genomic and phenotypic characterization. *PLoS ONE* **4**, e4982 (2009).
- Wyres, K. L., Lam, M. M. C. & Holt, K. E. Population genomics of *Klebsiella pneumoniae*. *Nat. Rev. Microbiol.* **18**, 344–359 (2020).
- Walker, K. A. et al. A *Klebsiella pneumoniae* regulatory mutant has reduced capsule expression but retains hypermucoviscosity. *MBio* **10**, e00089–19 (2019).
- Walker, K. A., Treat, L. P., Sepúlveda, V. E. & Miller, V. L. The small protein RmpD drives hypermucoviscosity in *Klebsiella pneumoniae*. *MBio* **11**, e01750–20 (2020).
- Holt, K. E. et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl Acad. Sci. USA* **112**, E3574–81 (2015).
- Gu, D. et al. A fatal outbreak of ST11 carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in a Chinese hospital: a molecular epidemiological study. *Lancet Infect. Dis.* **18**, 37–46 (2018).
- Xu, M. et al. High prevalence of KPC-2-producing hypervirulent *Klebsiella pneumoniae* causing meningitis in Eastern China. *Infect. Drug Resist.* **12**, 641–653 (2019).
- Dong, N. et al. Genome analysis of clinical multilocus sequence Type 11 *Klebsiella pneumoniae* from China. *Microb. Genomics* **4**, e000149 (2018).
- Wong, M. H. Y. et al. Emergence of carbapenem-resistant hypervirulent *Klebsiella pneumoniae*. *Lancet Infect. Dis.* **18**, 24 (2018).
- Yao, H., Qin, S., Chen, S., Shen, J. & Du, X.-D. Emergence of carbapenem-resistant hypervirulent *Klebsiella pneumoniae*. *Lancet Infect. Dis.* **18**, 25 (2018).
- Ørskov, I. D. A. & Fife-Asbury, M. A. New *Klebsiella* capsular antigen, K82, and the deletion of five of those previously assigned. *Int. J. Syst. Bacteriol.* **27**, 386–387 (1977).
- Trautmann, M. et al. O-antigen seroepidemiology of *Klebsiella* clinical isolates and implications for immunoprophylaxis of *Klebsiella* infections. *Clin. Diagn. Lab. Immunol.* **4**, 550–555 (1997).
- Elhani, D. et al. Molecular epidemiology of extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* strains in a university hospital in Tunisia, Tunisia, 1999–2005. *Clin. Microbiol. Infect.* **16**, 157–164 (2010).
- Chen, L. et al. Carbapenemase-producing *Klebsiella pneumoniae*: molecular and genetic decoding. *Trends Microbiol.* **22**, 686–696 (2014).
- Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. & Brisse, S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J. Clin. Microbiol.* **43**, 4178–4182 (2005).
- Wyres, K. L. & Holt, K. E. *Klebsiella pneumoniae* population genomics and antimicrobial-resistant clones. *Trend Microbiol.* **24**, 944–956 (2016).
- Long, S. W. et al. Population genomic analysis of 1,777 extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* isolates, Houston, Texas: unexpected abundance of clonal group 307. *MBio* **8**, e00489–17 (2017).
- Potter, R. F. et al. Population structure, antibiotic resistance, and uropathogenicity of *Klebsiella variicola*. *MBio* **9**, e02481–18 (2018).
- David, S. et al. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat. Microbiol.* <https://doi.org/10.1038/s41564-019-0492-8> (2019).
- Lam, M. M. C. et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in *Klebsiella pneumoniae* populations. *Microb. Genom.* **9**, e000196 (2018).

35. Lam, M. C. C. et al. Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in *Klebsiella pneumoniae*. *Genome Med.* **10**, 77 (2018).
36. Wick, R. R., Heinz, E., Holt, K. E. & Wyres, K. L. Kaptive Web: user-friendly capsule and lipopolysaccharide serotype prediction for *Klebsiella* genomes. *J. Clin. Microbiol.* **56**, e00197–18 (2018).
37. Martínez-Romero, E. et al. Genome misclassification of *Klebsiella variicola* and *Klebsiella quasipneumoniae* isolated from plants, animals and humans. *Salud Publica Mex.* **60**, 52–62 (2018).
38. Rodrigues, C. et al. Description of *Klebsiella africanensis* sp. nov., *Klebsiella variicola* subsp. *tropicalensis* subsp. nov. and *Klebsiella variicola* subsp. *variicola* subsp. nov. *Res. Microbiol.* **S0923-2508**, 30019–1 (2019).
39. Long, S. W. et al. Whole-genome sequencing of a human clinical isolate of the novel species *Klebsiella quasivariicola* sp. nov. *Genome Announc.* **5**, e01057–17 (2017).
40. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
41. Fajardo-Lubia n, A., Ben Zakour, N. L., Agyekum, A., Qi, Q. & Iredell, J. R. Host adaptation and convergent evolution increases antibiotic resistance without loss of virulence in a major human pathogen. *PLoS Pathog.* **15**, e1007218 (2019).
42. Wong, J. L. C. et al. OmpK36-mediated carbapenem resistance attenuates ST258 *Klebsiella pneumoniae* in vivo. *Nat. Commun.* **10**, 3957 (2019).
43. Hauck, C. et al. Spectrum of excess mortality due to carbapenem-resistant *Klebsiella pneumoniae* infections. *Clin. Microbiol. Infect.* **22**, 513–519 (2016).
44. Opoku-Temeng, C., Kobayashi, S. D. & DeLeo, F. R. *Klebsiella pneumoniae* capsule polysaccharide as a target for therapeutics and vaccines. *Comput. Struct. Biotechnol. J.* **17**, 1360–1366 (2019).
45. Venturini, C. et al. Fine capsule variation affects bacteriophage susceptibility in *Klebsiella pneumoniae* ST258. *FASEB J.* **34**, 10801–10817 (2020).
46. Pan, Y.-J. et al. Identification of three podoviruses infecting *Klebsiella* encoding capsule depolymerases that digest specific capsular types. *Microb. Biotechnol.* **12**, 472–486 (2019).
47. de Sousa, J. A. M., Buffet, A., Haudiquet, M., Rocha, E. P. C. & Rendueles, O. Modular prophage interactions driven by capsule serotype select for capsule loss under phage predation. *ISME J.* **14**, 2980–2996 (2020).
48. Arena, F. et al. Population structure of KPC carbapenemase-producing *Klebsiella pneumoniae* in a long-term acute-care rehabilitation facility: identification of a new lineage of clonal group 101, associated with local hyperendemicity. *Microb. Genom.* **6**, e000308 (2020).
49. Ferrari, C. et al. Multiple *Klebsiella pneumoniae* KPC clones contribute to an extended hospital outbreak. *Front. Microbiol.* **10**, 2767 (2019).
50. Magiorakos, A.-P. et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.* **18**, 268–281 (2012).
51. Wyres, K. L. et al. Genomic surveillance for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from South and Southeast Asia. *Genome Med.* **12**, 11 (2020).
52. Heinz, E., Brindle, R., Morgan-McCalla, A., Peters, K., & Thomson, N. R. Caribbean multi-centre study of *Klebsiella pneumoniae*: whole-genome sequencing, antimicrobial resistance and virulence factors. *Microb. Genom.* **5**, e000266 (2019).
53. Musicha, P. et al. Genomic analysis of *Klebsiella pneumoniae* isolates from Malawi reveals acquisition of multiple ESBL determinants across diverse lineages. *J. Antimicrob. Chemother.* **74**, 1223–1232 (2019).
54. Alvarez-Uria, G., Gandra, S., Mandal, S. & Laxminarayan, R. Global forecast of antimicrobial resistance in invasive isolates of *Escherichia coli* and *Klebsiella pneumoniae*. *Int. J. Infect. Dis.* **68**, 50–53 (2018).
55. Brolund, A. et al. Worsening epidemiological situation of carbapenemase-producing Enterobacteriaceae in Europe, assessment by national experts from 37 countries, July 2018. *Eurosurveillance* **24**, 1900123 (2019).
56. Coombs G. et al. on behalf of the Australian Group on Antimicrobial Resistance and Australian Commission on Safety and Quality in Health Care. Australian Group on Antimicrobial Resistance Sepsis Outcomes Programs: 2019 Report. Sydney: ACSQHC; 2021.
57. Kidd, T. J. et al. Molecular mechanisms and virulence of colistin-resistant *Klebsiella pneumoniae*. *Eur. Respir. J.* **48**, PA2625 (2016).
58. Cannatelli, A. et al. MgrB inactivation is a common mechanism of colistin resistance in KPC-producing *Klebsiella pneumoniae* of clinical origin. *Antimicrob. Agents Chemother.* **58**, 5696 LP–5703 (2014).
59. Cannatelli, A. et al. In vivo evolution to colistin resistance by PmrB sensor kinase mutation in KPC-producing *Klebsiella pneumoniae* is associated with low-dosage colistin treatment. *Antimicrob. Agents Chemother.* **58**, 4399–4403 (2014).
60. Yu, W.-L., Lee, M.-F., Tang, H.-J., Chang, M.-C. & Chuang, Y.-C. Low prevalence of *rmpA* and high tendency of *rmpA* mutation correspond to low virulence of extended spectrum  $\beta$ -lactamase-producing *Klebsiella pneumoniae* isolates. *Virulence* **6**, 162–172 (2015).
61. Chen, L. & Kreiswirth, B. N. Convergence of carbapenem-resistance and hypervirulence in *Klebsiella pneumoniae*. *Lancet Infect. Dis.* **18**, 2–3 (2018).
62. Lam, M. M. C. et al. Convergence of virulence and multidrug resistance in a single plasmid vector in multidrug-resistant *Klebsiella pneumoniae* ST15. *J. Antimicrob. Chemother.* <https://doi.org/10.1093/jac/dkz028> (2019).
63. Chen, Y. et al. Preterm infants harbour diverse *Klebsiella* populations, including atypical species that encode and produce an array of antimicrobial resistance- and virulence-associated factors. *Microb. Genomics* **6**, e000377 (2020).
64. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
65. Agyekum, A. et al. Predictability of phenotype in relation to common  $\beta$ -lactam resistance mechanisms in *Escherichia coli* and *Klebsiella pneumoniae*. *J. Clin. Microbiol.* **54**, 1243–1250 (2016).
66. Ginn, A. N. et al. Limited diversity in the gene pool allows prediction of third-generation cephalosporin and aminoglycoside resistance in *Escherichia coli* and *Klebsiella pneumoniae*. *Int. J. Antimicrob. Agents* **42**, 19–26 (2013).
67. Ginn, A. N. et al. Prediction of major antibiotic resistance in *Escherichia coli* and *Klebsiella pneumoniae* in Singapore, USA and China using a limited set of gene targets. *Int. J. Antimicrob. Agents* **43**, 563–565 (2014).
68. Stoesser, N. et al. Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data. *J. Antimicrob. Chemother.* **68**, 2234–2244 (2013).
69. Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M. & Lund, O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.* **71**, 2484–2488 (2016).
70. Bachman, M. A. et al. *Klebsiella pneumoniae* yersiniabactin promotes respiratory tract infection through evasion of lipocalin 2. *Infect. Immun.* **79**, 3309–3316 (2011).
71. Nougayrède, J. P. et al. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **313**, 848–851 (2006).
72. Russo, T. A., Olson, R., Macdonald, U., Beanan, J. & Davidson, B. A. Aerobactin, but not yersiniabactin, salmochelin, or enterobactin, enables the growth/survival of hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* ex vivo and in vivo. *Infect. Immun.* **83**, 3325–3333 (2015).
73. Russo, T. A. et al. Identification of biomarkers for the differentiation of hypervirulent *Klebsiella pneumoniae* from classical *K. pneumoniae*. *J. Clin. Microbiol.* **56**, e00776-18 (2018).
74. Konopka, K., Bindereif, A. & Neilands, J. B. Aerobactin-mediated utilization of transferrin iron. *Biochemistry* **21**, 6503–6508 (1982).
75. Lam, M. M. C. et al. Population genomics of hypervirulent *Klebsiella pneumoniae* clonal-group 23 reveals early emergence and rapid global dissemination. *Nat. Commun.* **9**, 2703 (2018).
76. Drlica, K. & Zhao, X. DNA gyrase, topoisomerase IV, and the 4-quinolones. *Microbiol. Mol. Biol. Rev.* **61**, 377–392 (1997).
77. Brisse, S. et al. *wzi* Gene sequencing, a rapid method for determination of capsular type for *Klebsiella* strains. *J. Clin. Microbiol.* **51**, 4073–4078 (2013).
78. Wick, R. R., Judd, L. M., Gorrie, C. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
79. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
80. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
81. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
82. Lu, J., Bretwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ. Comput. Sci.* **3**, e104 (2017).
83. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
84. Lam, M. M. C. et al. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. Kleborate v2.0.0 <https://doi.org/10.5281/zenodo.4923015> (2020).
85. Bialek-davenet, S. et al. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. *Emerg. Infect. Dis.* **20**, 1812–1820 (2014).
86. Neubauer, S. et al. A Genotype-phenotype correlation study of SHV  $\beta$ -lactamases offers new insight into SHV resistance profiles. *Antimicrob. Agents Chemother.* **64**, e02293–19 (2020).

### Acknowledgements

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

(<http://pathogen.watch>). This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation (OPP1175797). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Genetic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. K.E.H is also supported by the Viertel Charitable Foundation of Australia (Senior Medical Research Fellowship). K.L.W is supported by the National Health and Medical Research Council of Australia (Investigator Grant APP1176192).

### Author contributions

Study design: K.E.H. Data analysis: M.M.C.L., K.L.W. and K.E.H. Code development: R.R.W., K.E.H., S.C.W., L.T.C., and K.L.W. Manuscript writing: M.M.C.L., K.L.W., and K.E.H. All authors contributed to manuscript editing.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-24448-3>.

**Correspondence** and requests for materials should be addressed to M.M.C.L.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer review reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021