BMC Microbiology

RESEARCH

Open Access

Genomic typing, antimicrobial resistance gene, virulence factor and plasmid replicon database for the important pathogenic bacteria *Klebsiella pneumoniae*

Andrey Shelenkov[1*], Anna Slavokhotova[1], Yulia Mikhaylova[1] and Vasiliy Akimkin[1]

Abstract

Background The infections of bacterial origin represent a significant problem to the public healthcare worldwide both in clinical and community settings. Recent decade was marked by limiting treatment options for bacterial infections due to growing antimicrobial resistance (AMR) acquired and transferred by various bacterial species, especially the ones causing healthcare-associated infections, which has become a dangerous issue noticed by the World Health Organization. Numerous reports shown that the spread of AMR is often driven by several species-specific lineages usually called the 'global clones of high risk'. Thus, it is essential to track the isolates belonging to such clones and investigate the mechanisms of their pathogenicity and AMR acquisition. Currently, the whole genome-based analysis is more and more often used for these purposes, including the epidemiological surveillance and analysis of mobile elements involved in resistance transfer. However, in spite of the exponential growth of available bacterial genomes, their representation usually lack uniformity and availability of supporting metadata, which creates a bottleneck for such investigations.

Description In this database, we provide the results of a thorough genomic analysis of 61,857 genomes of a highly dangerous bacterial pathogen *Klebsiella pneumoniae*. Important isolate typing information including multilocus sequence typing (MLST) types (STs), assignment of the isolates to known global clones, capsular (KL) and lipooligosaccharide (O) types, the presence of CRISPR-Cas systems, and cgMLST profiles are given, and the information regarding the presence of AMR, virulence genes and plasmid replicons within the genomes is provided.

Conclusion This database is freely available under CC BY-NC-SA at https://doi.org/10.5281/zenodo.11069018. The database will facilitate selection of the proper reference isolate sets for any types of genome-based investigations. It will be helpful for investigations in the field of *K. pneumoniae* genomic epidemiology, as well as antimicrobial resistance analysis and the development of prevention measures against this important pathogen.

Keywords *Klebsiella pneumoniae*, Healthcare-associated infections, Genomic epidemiology, Whole genome sequencing, CgMLST, Global clones, Antibiotic resistance, Virulence factors

Background

Infections of bacterial origin represent one of the most serious problems in global healthcare. The treatment of such infections is complicated due to the spread of antimicrobial drug resistance (AMR) in clinical pathogenic bacteria, which leads to a limited set of

*Correspondence:
Andrey Shelenkov
fallandar@gmail.com
[1] Central Research Institute of Epidemiology, Novogireevskaya Str., 3a, Moscow 111123, Russia

available treatment options [1]. However, AMR can also be acquired by the non-hospital bacterial populations, thus making this problem global and becoming an issue in community settings [2].

The spread of AMR within the populations of particular bacterial species is often facilitated by several groups or lineages called 'global clones' or 'clonal groups'. An important role of such clones has been demonstrated for some of the most prominent and widespread nosocomial pathogens like *Klebsiella pneumoniae* [3, 4], *Acinetobacter baumannii* [5] and *Pseudomonas aeruginosa* [6]. Therefore, epidemiological surveillance of multidrug-resistant (MDR) bacteria and developing effective measures of prevention against their spreading should involve checking if particular isolates belong to such global clones.

Isolate typing and assignment to a particular clonal group, or global clone, can be determined on the basis of several characteristics obtained by molecular biology techniques, but now the whole genome sequencing (WGS) is becoming a gold standard for this and many other investigations due to dramatic increase in the amount of data it produces and its already recognized labor- and cost-effectiveness [7, 8]. Currently, several hundred thousand genomes of pathogenic bacteria are available in various public databases, including NCBI Genbank (https://www.ncbi.nlm.nih.gov/genbank/), and increasing availability of WGS will facilitate further growth in amount of such data.

The fraction of genomes available in public databases for particular bacterial species is determined by its significance to public healthcare and the incidence of infections caused by this species, and *K. pneumoniae* is responsible for a large share of nosocomial and community-acquired infections worldwide [9]. The World Health Organization (WHO) listed carbapenem-resistant *K. pneumoniae* in the group of the critical pathogens having the highest priority in new antibiotic development [10].

At present, more than 70,000 draft *K. pneumoniae* genomes are available at NCBI (https://www.ncbi.nlm.nih.gov/datasets/taxonomy/573/, accessed on 8 September 2024). However, the data provided there do not contain important typing information and other metadata, including AMR gene presence, which is supposed to be derived by users using third-party computational pipelines. Another well-known database PasteurMLST [11] contains typing information and epidemiological metadata for more than 40,000 genomes (https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_isolates, accessed on 8 September 2024). However, the extraction of AMR, virulence gene and plasmid information from this database is not straightforward, especially for the large number of isolates simultaneously. In addition, no mechanism is provided for quick download of multiple genomes satisfying several selection criteria at once.

In order to facilitate the rapid construction of *K. pneumoniae* genome subsets for particular purpose, we developed the database containing typing information, including clonal group determination, for the whole set of the *K. pneumoniae* isolates available at NCBI genome assembly database (https://www.ncbi.nlm.nih.gov/datasets/taxonomy/573/, accessed on 14 December 2023), which included 61,857 genomes. The data provided includes multilocus sequence typing (MLST)-based sequence types (STs), capsular polysaccharide (KL, or K) and lipooligosaccharide (O) types, as well as core genome MLST (cgMLST) profiles, and the information concerning the presence and type of CRISPR-Cas systems in *K. pneumoniae* isolates. The information regarding AMR gene presence, which can confer resistance to various classes of antibiotics, as well as the set of virulence factors and the presence of plasmid replicons, is also available.

The combination of several typing schemes, e.g., ST, KL and O types, was previously shown to provide significantly higher resolution than any of these schemes used alone, while cgMLST possessed the highest discrimination ability [12]. However, other specific genomic features can be used for classification purposes when deemed useful, for example, CRISPR profiles or even repeat frequency characteristics [13].

This database was used for selection of appropriate reference genomes for comparison purposes in our previous research [3, 14] and found to be very convenient for these purposes, so we decided making it publicly available for the researchers working on *K. pneumoniae* genomic analysis.

Construction and content

We retrieved 62,859 genomic sequences of *K. pneumoniae* from Genbank (https://www.ncbi.nlm.nih.gov/genbank/, accessed on 14 December 2023), for which the assembly level was defined as 'Complete Genome', 'Chromosome', or 'Scaffold'. Then we excluded 1002 isolates (about 1.6%) for which the MLST revealed inexact scheme or shown the incorrect species identification. The rest of the isolates, constituting 61,857 in total, were also checked to have higher than 95% pairwise average nucleotide identity, and no further isolates were excluded based on this criterion, which was recently confirmed to be reliable threshold for species differentiation [15].

MLST was performed using the Institut Pasteur database (https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef, accessed on 12 August 2024) using the typing scheme described in [16].

The evolution of the multidrug resistance in *K. pneumoniae* is largely driven by the plasmid-mediated acquisition of resistance genes, and several hospital outbreak clones, also known as 'global clones' or 'clonal groups', were shown to play an important role in this process [17, 18]. Thus, an assignment of the isolates to known global clones represents an important step in epidemiological surveillance and antimicrobial resistance spread investigation. In this database, such an assignment was based on MLST ST according to previously published data [4, 18].

The AMR genes were detected using Resfinder 4.6.0 software [19] (http://genepi.food.dtu.dk/resfinder, accessed on 10 August 2024, using the default parameters).

The detection of capsule polysaccharide loci (KL) and lipooligosaccharide loci (O) was made using Kaptive v. 2.0.9 [20] with the default parameters (last update of the databases was on 12 March 2024).

Searching for virulence factors in the *K. pneumoniae* genomes was performed using VFDB [21] (http://www.mgc.ac.cn/VFs/main.htm, accessed on 20 August 2024, using the default parameters).

Plasmid replicons were detected using PlasmidFinder 2.1 with default parameters [22] (https://cge.food.dtu.dk/services/PlasmidFinder/, accessed on December 1, 2024).

The presence of CRISPR-Cas systems in the genomes under study was analyzed using CRISPRCasFinder [23] version 4.2.20 with the following parameters: '-fast -rcfowce -ccvRep -vicinity 1200 –cas -useProkka'. The variants of CRISPR-Cas systems (I-E vs. I-E*) were determined according to the literature data [24].

The cgMLST profiles were built using MentaList software [25] (https://github.com/WGS-TB/MentaLiST, version 0.2.4, default parameters, accessed on 21 August 2024) using the scheme with 2358 loci obtained from cgmlst.org (https://www.cgmlst.org/ncs/schema/schema/2187931/, accessed on 12 August 2024).

Additional processing of the data, formatting output files and preparation of the tables for the database were performed using the computational pipeline, which was developed by us earlier and was already applied to several investigations [26, 27].

The database contains four tables provided in various formats: xlsx (all except cgMLST), tab-delimited txt and pdf (for summary table only). The files in xlsx format can be further processed by users in table processing software, as described previously [28]. Text format (txt) can be used for computational processing by various bioinformatics instruments, while pdf format presents the summary containing the most important typing results, which are presented in ready-to-read form.

The tables, which will be further described below, include the following:

- A summary table (table_summary) containing typing information for all isolates, such as MLST ST, KL- and O-types, clonal groups, as well as the presence and type of CRISPR-Cas systems in the genomes
- An AMR gene table (table_amr) containing the information on the presence of the genes previously reported to confer the AMR to various classes of antimicrobial drugs in the genomes of all isolates
- A virulence gene table (table_vfdb) containing the information on the presence of genes encoding virulence factors
- A plasmid replicon table (table_plasmid) including the data on the plasmid replicons revealed in the genomes of the isolates
- A table containing cgMLST profiles (table_cgmlst) for the isolates, which can be applied to more detailed comparison and provides better resolution than MLST

Utility and discussion

Table structure and content description

The format and exemplary data for the summary table are presented in Table 1.

The first column contains the assembly code from Genbank, which uniquely identifies a particular *K. pneumoniae* genome assembly and is used in all tables.

'CG' stands for 'clonal group' and indicates the assignment of a particular isolate to known clonal groups. Such an assignment is based on the ST according to the previously published data [4, 18]. If an isolate was not assigned

Table 1 Exemplary data and column information for the typing summary table

Assembly code	Clonal group	MLST ST	KL type	O type	CRISPR-Cas
GCA_000009885.1	CG23	23	KL1	O1/O2v2	I-E*
GCA_000255975.2	NA[1]	48	KL124	O1/O2v1	NF[3]
GCA_005508835.1	NA	ND[2]	KL15	O4	UNKN

[1] *'NA'* (not available) indicates that the isolate did not belong to any known clonal group

[2] *'ND'* (not determined) indicates that ST could not be determined

[3] *'NF'* (not found) – CRISPR-Cas system was not found in the genome

to any CG, this column contains 'NA' (not available) designation. The composition of CG is given in Table 2 for reference purposes.

Third column contains an ST defined as a combination of seven loci (*gapA*, *infB*, *mdh*, *pgi*, *phoE*, *rpoB* and *tonB* genes) of a typing scheme [16]. Each variant of a particular locus is numbered sequentially, and the unique combination of seven locus variants constitutes a ST, to which its own number is assigned. For example, the combination of *gapA_2, infB_1, mdh_1, pgi_1, phoE_9, rpoB_4* and *tonB_12* alleles defines the ST23. Locus variants and the corresponding STs can be found in the Institut Pasteur database (https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef, accessed on 12 August 2024). 'ND' in this column indicates that ST was not determined due to either low sequencing quality, insufficient genome region coverage or the presence of a novel MLST allele, or novel allele combination, not yet uploaded to the databases.

Table 2 Composition of *K. pneumoniae* clonal groups based on MLST ST

Clonal group	STs
CG15	14, 15, 709
CG20	16, 17, 20, 22, 336, 1123
CG23	23, 26, 57, 163, 218, 260, 887, 2044
CG25	25
CG29	29, 711, 1109, 1271
CG34	34, 228, 592, 1444, 1454, 1521, 2141
CG35	35, 466, 1948
CG36	36, 268, 298, 433, 564, 1272, 2130
CG37	37, 177, 256, 309, 896, 1198, 1507, 1779
CG43	43, 101, 2017
CG45	45, 485
CG48	48
CG65	25, 65, 375
CG66	66, 2039, 2058
CG86	86, 373, 3994
CG101	101, 1685, 2016, 2017, 2502
CG111	111, 692
CG133	133, 420, 2127
CG147	147, 273, 392, 1709
CG152	105, 152
CG230	224, 230, 1307, 1583
CG231	231, 1190
CG253	253, 1138, 2116
CG258	11, 258, 340, 379, 395, 418, 437, 512, 833, 855, 1084, 1199
CG322	322, 491, 1377
CG380	380, 679
CG490	76, 199, 490, 495, 1263
CG540	4, 6, 504, 540, 1013
CG661	237, 661, 2113

KL- and O-type show the typing variants based on the corresponding sets of genes, respectively. Capsular polysaccharide is an essential factor, which influences bacterial virulence and susceptibility to different bacteriophages, and thus becomes an important epidemiological marker [29]. The typing of a capsular polysaccharide (*cps*) gene cluster in *K. pneumoniae* was based on the *wzi* or *wzc* gene sequences, while O-locus typing was defined by sequence identity in the conserved *wzm* and *wzt* genes. Each distinct gene cluster found between the flanking genes is given a unique code identifying the cluster type, and these data can also be found in public databases [20].

The final column indicates the presence and type of a CRISPR-Cas system in the isolate. Clustered regularly interspaced short palindromic repeat (CRISPR) arrays and CRISPR-associated genes (*cas*) function as a variable genetic element and form bacterial adaptive immune systems. CRISPR-Cas systems are currently divided into six major types (I-VI) and several subtypes (A-I, K, U) based on a combination of various genomic and structural analysis [30, 31]. 'NF' in this column indicates that CRISPR-Cas system was not revealed in a particular isolate genome, 'UNKN' shows the detected presence of an incomplete or untypeable system, and in other cases a system type is shown.

Another part of the database includes information regarding the presence of various genes known to confer AMR, which has been revealed in the *K. pneumoniae* genomes under investigation. The first column shows the assembly code, which is identical to the one from the typing summary table; the second column gives the number of AMR genes revealed in a particular isolate, while other columns describe the presence of a particular gene from the first row and its sequence similarity level with the corresponding allele from the Resfinder database. The absence of a gene is marked with a dot to increase the readability.

An example is provided in Table 3. Only a few genes are presented in this example.

It should be noted that the presence of a particular gene known to provide a resistance to some antimicrobial drug per se does not confirm the phenotypic resistance to this drug since this gene, for example, might not be expressed at all or expressed on the insufficient level [12, 32]. However, this information is essential for the estimation of the AMR repertoire and spread within bacterial population of interest.

The format of the virulence gene table is the same as the one for AMR gene table except for the set of genes

Table 3 Exemplary data and column information for antimicrobial resistance gene table

Assembly code	NUM_FOUND	*oqxA*	*oqxB*	…	*blaTEM-1B*	*catA1*
GCA_000016305.1	19	100.00	100.00	…	99.8	100.00
GCA_000219965.2	13	100.00	100.00	…	.	.

included. The table containing plasmid replicons also has the same format.

The information presented in AMR and virulence gene tables can be useful for the comparative analysis of pathogenicity within particular groups of the isolates. Genetically and epidemiologically close isolates can possess different AMR and virulence gene repertoire, and thus this information is essential for the selection of proper reference sets. The co-presence of certain plasmid replicons, like IncFIB, IncFII or IncR, and AMR or virulence genes could possibly indicate the plasmid origin of such genes, but additional investigations are always needed to verify such a hypothesis.

The fourth part of the database includes cgMLST profiles for all isolates. The cgMLST typing scheme is similar to MLST in the sense that it enumerates the existing gene alleles and uses their combination as a profile, but the striking difference is that cgMLST relies on the set of all conservative genes present within particular species (usually, the genes appeared in more than 90% of known isolates). cgMLST scheme for *K. pneumoniae* includes 2358 loci, and the allele variants for these loci are available in a regularly updated database at cgmlst.org (https://www.cgmlst.org/ncs/schema/schema/2187931/, accessed on 12 August 2024).

cgMLST profiles can be used in the cluster analysis of a selected set of isolates for the estimation of their genomic similarity and revealing epidemiologically or clinically important groups. The threshold of 18 differences in cgMLST alleles was previously proposed to determine if two *K. pneumoniae* isolates belonged to a single strain or different strains [33], but less or more strict criteria can be used depending on the specific goal of investigations.

In this database, cgMLST profiles are provided in a table format. The first column contains the same assembly identifier as other tables, while the other columns show the numbers (or special symbols) representing the variants of genes appeared in a header row. Several special designations can appear in this column, such as, 'N'—indicates a novel allele variant not yet present in the database; '0?' indicates a locus missing in the assembly (due to the low quality of initial sample or sequencing problems); '- ' points an allele is partially covered; '+' represents multiple possible alleles, in which case the one with a highest probability is reported.

General descriptive statistics

In this section, some general descriptive statistics based on the database is provided. Previously it was shown that a Genbank set of genomes is not representative for the worldwide *K. pneumoniae* population since it was strongly biased towards MDR or other clinically relevant isolates from particular regions [34]. On the other hand, the descriptive statistics on the distribution of particular STs, CGs, AMR, virulence genes etc. can provide useful information for the purposes of reference set selection and making comparisons.

Below we will refer to any Genbank assembly record containing either complete or partial genome as an "isolate" for the sake of brevity, despite some different assemblies can in fact represent the same isolate, or some genomic records may contain only a part of the genome.

The summary of top three dominating characteristics in each feature category is given in Table 4.

Totally, 37.3% of the isolates belonged to known CGs, with CG258 accounting for 25.3% of all *K. pneumoniae* genomes from Genbank and 68% of *K. pneumoniae* isolates belonging to CGs, respectively. CG147 was the second largest CG, which included 6.2%/16.6% of all isolates/CG isolates. CG258 (including ST11 and ST258, among others), is known to be the dominant carbapenemase-producing *K. pneumoniae* clone worldwide, which raises global concerns in healthcare systems of many countries [35]. Thus, it is not surprising that it has the largest number of the sequenced isolates in Genbank, and two STs belonging to CG258 are also the most represented there with 15.5% for ST11 and 8% for ST258, respectively.

Table 4 Top three dominating representatives of the features investigated in *K. pneumoniae* database

Feature	Top three representatives
Clonal Group	NA, CG258, CG147
Sequence type	ST11, ST258, ST147
KL-type	KL64, KL107, KL2
O-type	O1/O2v1, O1/O2v2, O3b
AMR genes	*oqxAB, fosA6, sul1*
Virulence genes	*ompA, fur, acrAB*
Plasmid replicons	IncFII(pKP91), IncFIB(K), Col(pHAD28)
CRISPR-Cas system	NF, I-E, I-E*

ST147, a most prominent member of CG147, which is also recognized as a globally distributed highly resistant clone [36], accounted for 5.3% of the isolates.

In total, 2402 distinct STs were revealed in Genbank for *K. pneumoniae*, while ST could not be determined for 6.4% of the isolates due to low-quality sequencing and/or lack of coverage for the corresponding genomic regions. However, 1069 of these STs were represented by a single isolate only, and 389 top STs included ten or more isolates, totally enclosing 86% of the Genbank *K. pneumoniae* genomes.

Top three KL types included KL64 (12.9% of the isolates), KL107 (8.5%) and KL2 (5.5%). KL2 is considered to be predominantly associated with hypervirulent *K. pneumoniae* isolates [37]. The abundance of KL64 and KL107 is also not surprising since these types were often associated with CG258 and CG147, and ST11-KL64 and ST258-KL107 were even highlighted as a separate important multidrug-resistant clones [38, 39]. The diversity of KL types was high with 151 distinct types revealed.

The range of lipopolysaccharide types was less diverse with 12 distinct types, three of which were O1 variants. O1/O2v1 (37.3%) and O1/O2v2 (34.9%) together accounted for about 72% of the isolates, which corresponds to previous reports [40, 41], and O3b was revealed in about 8% of the isolates.

The number of AMR genes possessed by the isolates was in a range from 1 to 43 with a median equal to 13. This number was never equal to zero since intrinsic *blaSHV* gene variants and other intrinsic genes were also included. The absence of intrinsic AMR genes in some isolates could indicate sequencing or assembly problems.

The most abundant AMR genes were *oqxAB* encoding efflux pump conferring fluoroquinolone resistance. These genes were revealed in about 90% of the isolates, which corresponds to the recent clinical data [42]. High frequency of occurrence was also observed for *fosA6* (fosfomycin resistance, 61.6%) and *sul1* (sulfamethoxazole resistance, 44.4%). The first two factors, *oqxAB* and *fosA6*, were shown to be intrinsic in *K. pneumoniae* [43], and thus their abundance is not surprising. Top three acquired genes were the above-mentioned *sul1* (44.4%) and beta-lactamase-encoding $bla_{CTX\text{-}M\text{-}15}$ and $bla_{TEM\text{-}1B}$ accounting for about 38% each and representing the most widespread beta-lactamase genes except for intrinsic *blaSHV* group. The latter included more than 100 variants present in more than 97% of the isolates. In addition, carbapenemase-encoding $bla_{NDM\text{-}1}$ (about 10%) was the most abundant in the corresponding class. In total, 713 different AMR genes or gene variants were revealed, 324 of which were found in 10 or more isolates.

However, it should be noted again that the isolate distribution in Genbank is skewed towards MDR strains and thus cannot be considered representative for the whole population. On the other hand, these data provides useful insights into the procedure of relevant reference set selection for various *K. pneumoniae* investigations.

The number of virulence genes in the database varied from 27 to 131 with a median equal to 73. Among these, *ompA* (outer membrane protein, porin), *fur* (ferric uptake regulator) and *acrAB* (efflux pump) were revealed in virtually all isolates. Worth noting, 13.4% of the isolates carried *rmpA*, one of the major hypervirulence markers [44]. Totally, 428 distinct virulence genes were found, but only 137 of them occurred in a significant number (>=10) of the isolates.

The total number of the isolates including at least one plasmid replicon was 60,226 (97.3%). In turn, the number of replicons present varied from 0 to 14 with a median equal to 4. It should be noted that there is no one-to-one correspondence between the real plasmids and the replicons predicted since some plasmids can harbor more than one replicon [45], and the prediction can in some cases provide two or more version of the same replicon. The most widespread replicon was IncFII(pKP91) revealed in 33,483 isolates (54.1%) followed by IncFIB(K) (52.1%) and Col(pHAD28) (43.5%). IncR plasmids were previously shown to carry multiple AMR genes in clinical Enterobacterales, especially in *K. pneumoniae* [46], and this replicon was revealed in 32.1% of the isolates. Together with other sub-replicons (e.g., IncFII(pHN7A8), IncFII(pCRY) etc.) IncFII type replicons were found in 69.7% of the isolates, while IncFIB variants were revealed in 75% of the isolates. These data conforms well to previous reports showing that IncFIB and IncFII plasmids are widespread and strongly associated with AMR transfer in clinical *K. pneumoniae* [47]. In total, 162 different replicons were revealed, 128 of which were found more than 100 times.

CRISPR-Cas systems were found in about 29% (18,199) of the isolates, with I-E occurring slightly more often than I-E* (14.2% vs. 10.9% for all isolates and 48.2% vs. 37% for the isolates with CRISPR-Cas, respectively). These findings correspond to other recent reports, in which CRISPR-Cas prevalence in *K. pneumoniae* was estimated to be about 25% [24, 48]. Besides this, 2644 isolates included CRISPR-Cas systems which were incomplete or had undetectable type with alternative *cas* gene sequence. Interestingly, 67 isolates included two types of systems I-E and I-E* simultaneously, 15 of which belonged to ST1758-KL27-O4 type.

Another interesting fact was that about 92% of CG147 possessed CRISPR-Cas system, while CG258 virtually lacked it. The members of CG147 were previously shown to be abundant with CRISPR-Cas systems and even to possess both this system and antibiotic resistance-encoding plasmids simultaneously [49]. The absence of CRISPR-Cas in CG258 was also reported previously and linked to dissemination of IncF type plasmids in this group [50].

The full data describing the distribution of all the characteristics discussed above is presented in Table S1.

Applications and future development

The database is available for academic use under Creative Commons Attribution-Non Commercial-ShareAlike (CC BY-NC-SA) 4.0 International License. The updates are scheduled to be provided at least once a year.

In general, the database is intended to be used for several purposes. First, collecting various statistics on *K. pneumoniae* genomes available, for example, the prevalence of particular AMR genes in particular groups (STs, CGs) of the isolates, and searching for novel correlations, which can potentially be used in epidemiological surveillance. Such correlations might include the increased/decreased number of virulence/AMR genes for particular STs or CGs, or the prevalence of particular beta-lactamases in these groups. For example, the tendency of possessing higher virulence was previously shown for several STs and certain capsular serotypes [37].

Second, the database is useful for selecting reference genomes for comparison purposes. Bacterial genomic studies usually include making a comparison between the newly obtained isolates and the strains of the same ST or CG, or carrying the same set of AMR or virulence genes, which are contained in public databases. However, Genbank metadata in most cases do not provide this information, which can be obtained only from the corresponding publications, if any. PubMLST database includes such information, but the search can be based only on one ST/CG or AMR gene at a time, and the tools for genome selection and downloading are not straightforward to use and impose the limits on the number of genomes analyzed simultaneously. In our database, such a selection can be easily performed with Excel filters or Linux/Windows-based text manipulation software.

Third, the database can be used for studying the global prevalence of known global clones of high risk and searching for possible candidates to be included in a novel clonal group of the isolates.

The description of possible database applications to genomic epidemiology investigations, including the commands for UNIX-based systems, will be provided on the webpage of the database in a 'how_to_use.txt' file.

Conclusions

We have developed a database containing important genomic information for *K. pneumoniae*, a bacterial pathogen of a global high concern, based on Genbank genomic records. The data include typing information, including cgMLST profiles, as well as AMR and virulence gene presence information. Additionally, the data on plasmid replicon and CRISPR-Cas system presence is also provided. We believe that the database will be useful for genomic and epidemiological studies of *K. pneumoniae*, including selection of the proper reference isolate sets for any types of genome-based investigations. The precomputed data will be particularly useful for the researchers working in the promising field of genomic epidemiology of this important and highly dangerous pathogen.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12866-024-03720-8.

Supplementary Material 1: Table S1. General characteristics of *Klebsiella pneumoniae* genomes from Genbank.

Authors' contributions

Conceptualization, A.Sh., A.S and Y.M.; data curation, A. Sh.; formal analysis, A.Sh., Y.M., V.A.; funding acquisition, V.A.; investigation, A. Sh., A.S., Y.M., and V.A.; methodology, A.Sh. and Y.M.; project administration, V.A.; resources, A.Sh, A.S..; software, A.Sh.; supervision, V.A.; validation, A.Sh. and A.S.; visualization, A.Sh.; writing—original draft, A.Sh.; writing—review and editing, A.Sh., A.S., Y.M., V.A. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Ministry of Science and Higher Education of the Russian Federation within the framework of a grant in the form of a subsidy for the creation and development of the «World-class Genomic Research Center for Ensuring Biological Safety and Technological Independence under the Federal Scientific and Technical Program for the Development of Genetic Technologies», agreement No. 075–15-2019–1666. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Data availability

The datasets generated and/or analyzed during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.11069018. This page is accessible via any web browser and all tables can be downloaded there. The analysis and descriptive statistics data are included in this published article and its supplementary information files. The draft version of this manuscript is available as a preprint at https://doi.org/10.20944/preprints202409.1714.v1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 9 October 2024 Accepted: 19 December 2024
Published online: 06 January 2025

References

1. Salam MA, Al-Amin MY, Salam MT, Pawar JS, Akhter N, Rabaan AA, et al. Antimicrobial Resistance: A Growing Serious Threat for Global Public Health. Healthcare. 2023;11(13):1946.
2. Denissen J, Reyneke B, Waso-Reyneke M, Havenga B, Barnard T, Khan S, et al. Prevalence of ESKAPE pathogens in the environment: Antibiotic resistance status, community-acquired infection and risk to human health. Int J Hyg Environ Health. 2022;244:114006.
3. Shaidullina ER, Schwabe M, Rohde T, Shapovalova VV, Dyachkova MS, Matsvay AD, et al. Genomic analysis of the international high-risk clonal lineage Klebsiella pneumoniae sequence type 395. Genome Med. 2023;15(1):9.
4. Arcari G, Carattoli A. Global spread and evolutionary convergence of multidrug-resistant and hypervirulent Klebsiella pneumoniae high-risk clones. Pathog Glob Health. 2023;117(4):328–41.
5. Zhang X, Li F, Awan F, Jiang H, Zeng Z, Lv W. Molecular Epidemiology and Clone Transmission of Carbapenem-Resistant Acinetobacter baumannii in ICU Rooms. Front Cell Infect Microbiol. 2021;11:633817.
6. Zurita J, Sevillano G, Solis MB, Paz YMA, Alves BR, Changuan J, et al. Pseudomonas aeruginosa epidemic high-risk clones and their association with multidrug-resistant. J Glob Antimicrob Resist. 2024;38:332–8.
7. Atxaerandio-Landa A, Arrieta-Gisasola A, Laorden L, Bikandi J, Garaizar J, Martinez-Malaxetxebarria I, et al. A Practical Bioinformatics Workflow for Routine Analysis of Bacterial WGS Data. Microorganisms. 2022;10(12):2364.
8. Price V, Ngwira LG, Lewis JM, Baker KS, Peacock SJ, Jauneikaite E, et al. A systematic review of economic evaluations of whole-genome sequencing for the surveillance of bacterial pathogens. Microb Genom. 2023;9(2):mgen000947.
9. Lin XC, Li CL, Zhang SY, Yang XF, Jiang M. The Global and Regional Prevalence of Hospital-Acquired Carbapenem-Resistant Klebsiella pneumoniae Infection: A Systematic Review and Meta-analysis. Open Forum Infect Dis. 2024;11(2):ofad649.
10. WHO. WHO Bacterial Priority Pathogens List,. bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance. Geneva: World Health Organization; 2024. p. 2024.
11. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.
12. Shelenkov A, Mikhaylova Y, Yanushevich Y, Samoilov A, Petrova L, Fomina V, et al. Molecular Typing, Characterization of Antimicrobial Resistance, Virulence Profiling and Analysis of Whole-Genome Sequence of Clinical Klebsiella pneumoniae Isolates. Antibiotics (Basel). 2020;9(5):261.
13. Shelenkov A, Skryabin K, Korotkov E. Search and classification of potential minisatellite sequences from bacterial genomes. DNA Res. 2006;13(3):89–102.
14. Shelenkov A, Mikhaylova Y, Voskanyan S, Egorova A, Akimkin V. Whole-Genome Sequencing Revealed the Fusion Plasmids Capable of Transmission and Acquisition of Both Antimicrobial Resistance and Hypervirulence Determinants in Multidrug-Resistant Klebsiella pneumoniae Isolates. Microorganisms. 2023;11(5):1314.
15. Rodriguez RL, Conrad RE, Viver T, Feistel DJ, Lindner BG, Venter SN, et al. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio. 2024;15(1):e0269623.
16. Diancourt L, Passet V, Verhoef J, Grimont PA, Brisse S. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. J Clin Microbiol. 2005;43(8):4178–82.
17. Peirano G, Chen L, Kreiswirth BN, Pitout JDD. Emerging Antimicrobial-Resistant High-Risk Klebsiella pneumoniae Clones ST307 and ST147. Antimicrob Agents Chemother. 2020;64(10):e01148-20.
18. Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. PLoS Genet. 2019;15(4):e1008114.
19. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother. 2020;75(12):3491–500.
20. Lam MMC, Wick RR, Judd LM, Holt KE, Wyres KL. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species complex. Microb Genom. 2022;8(3):000800.
21. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res. 2022;50(D1):D912–7.
22. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895–903.
23. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Neron B, et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. 2018;46(W1):W246–51.
24. Kannadasan AB, Sumantran VN, Vaidyanathan R. A Global Comprehensive Study of the Distribution of Type I-E and Type I-E* CRISPR-Cas Systems in Klebsiella pneumoniae. Indian J Community Med. 2023;48(4):567–72.
25. Feijao P, Yao HT, Fornika D, Gardy J, Hsiao W, Chauve C, et al. MentaLiST - A fast MLST caller for large MLST schemes. Microb Genom. 2018;4(2):e000146.
26. Egorova A, Mikhaylova Y, Saenko S, Tyumentseva M, Tyumentsev A, Karbyshev K, et al. Comparative Whole-Genome Analysis of Russian Foodborne Multidrug-Resistant Salmonella Infantis Isolates. Microorganisms. 2021;10(1):89.
27. Shelenkov A, Petrova L, Fomina V, Zamyatin M, Mikhaylova Y, Akimkin V. Multidrug-Resistant Proteus mirabilis Strain with Cointegrate Plasmid. Microorganisms. 2020;8(11):1775.
28. Shelenkov A, Mikhaylova Y, Akimkin V. Genomic Epidemiology Dataset for the Important Nosocomial Pathogenic Bacterium Acinetobacter baumannii. Data. 2024;9(2):22.
29. Arbatsky NP, Shneider MM, Dmitrenok AS, Popova AV, Shagin DA, Shelenkov AA, et al. Structure and gene cluster of the K125 capsular polysaccharide from Acinetobacter baumannii MAR13-1452. Int J Biol Macromol. 2018;117:1195–9.
30. Makarova KS, Koonin EV. Annotation and Classification of CRISPR-Cas Systems. Methods Mol Biol. 2015;1311:47–75.
31. Hryhorowicz M, Lipinski D, Zeyland J. Evolution of CRISPR/Cas Systems for Precise Genome Editing. Int J Mol Sci. 2023;24(18):14233.
32. Stasiak M, Mackiw E, Kowalska J, Kucharek K, Postupolski J. Silent Genes: Antimicrobial Resistance and Antibiotic Production. Pol J Microbiol. 2021;70(4):421–9.
33. Schurch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin Microbiol Infect. 2018;24(4):350–4.
34. Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat Commun. 2021;12(1):4188.
35. Liu C, Wu Y, Fang Y, Sang Z, Huang L, Dong N, et al. Emergence of an ST1326 (CG258) Multi-Drug Resistant Klebsiella pneumoniae Co-harboring mcr-8.2, ESBL Genes, and the Resistance-Nodulation-Division Efflux Pump Gene Cluster tmexCD1-toprJ1 in China. Front Microbiol. 2022;13:800993.
36. Rodrigues C, Desai S, Passet V, Gajjar D, Brisse S. Genomic evolution of the globally disseminated multidrug-resistant Klebsiella pneumoniae clonal group 147. Microb Genom. 2022;8(1):000737.
37. Kocsis B. Hypervirulent Klebsiella pneumoniae: An update on epidemiology, detection and antibiotic resistance. Acta Microbiol Immunol Hung. 2023;70(4):278–87.
38. Wang J, Feng Y, Zong Z. The Origins of ST11 KL64 Klebsiella pneumoniae: a Genome-Based Study. Microbiol Spectr. 2023;11(2):e0416522.
39. Wu JW, Quyen TLT, Hsieh YC, Chen YY, Wu LT, Pan YJ. Investigation of carbapenem-resistant Klebsiella pneumoniae in Taiwan revealed strains co-harbouring bla(NDM) and bla(OXA-48-like) and a novel plasmid co-carrying bla(NDM-1) and bla(OXA-181). Int J Antimicrob Agents. 2023;62(5):106964.
40. Choi M, Hegerle N, Nkeze J, Sen S, Jamindar S, Nasrin S, et al. The Diversity of Lipopolysaccharide (O) and Capsular Polysaccharide (K) Antigens

of Invasive Klebsiella pneumoniae in a Multi-Country Collection. Front Microbiol. 2020;11:1249.

41. Mills RO, Dadzie I, Le-Viet T, Baker DJ, Addy HPK, Akwetey SA, et al. Genomic diversity and antimicrobial resistance in clinical Klebsiella pneumoniae isolates from tertiary hospitals in Southern Ghana. J Antimicrob Chemother. 2024;79(7):1529–39.
42. Mihailovskaya VS, Selivanova PA, Kuznetsova MV. Prevalence of qacEΔ1, qacE, oqxA, oqxB, acrA, cepA and zitB genes among multidrug-resistant Klebsiella pneumoniae isolated in a cardiac hospital. J Microbiol Epidemiol Immunobiol. 2024;101(4):502–11.
43. Bernardini A, Cuesta T, Tomas A, Bengoechea JA, Martinez JL, Sanchez MB. The intrinsic resistome of Klebsiella pneumoniae. Int J Antimicrob Agents. 2019;53(1):29–33.
44. Russo TA, Carlino-MacDonald U, Drayer ZJ, Davies CJ, Alvarado CL, Hutson A, et al. Deciphering the relative importance of genetic elements in hypervirulent Klebsiella pneumoniae to guide countermeasure development. EBioMedicine. 2024;107:105302.
45. Wang X, Zhao J, Ji F, Chang H, Qin J, Zhang C, et al. Multiple-Replicon Resistance Plasmids of Klebsiella Mediate Extensive Dissemination of Antimicrobial Genes. Front Microbiol. 2021;12:754931.
46. Qian C, Zhu X, Lu J, Shen K, Chen Q, Zhou W, et al. Characterization of an IncR Plasmid with Two Copies of ISCR-Linked qnrB6 from ST968 Klebsiella pneumoniae. Int J Genomics. 2020;2020:3484328.
47. Pankok F, Taudien S, Dekker D, Thye T, Oppong K, Wiafe Akenten C, et al. Epidemiology of Plasmids in Escherichia coli and Klebsiella pneumoniae with Acquired Extended Spectrum Beta-Lactamase Genes Isolated from Chronic Wounds in Ghana. Antibiotics (Basel). 2022;11(5):689.
48. Alkompoz AK, Hamed SM, Zaid ASA, Almangour TA, Al-Agamy MH, Aboshanab KM. Correlation of CRISPR/Cas and Antimicrobial Resistance in Klebsiella pneumoniae Clinical Isolates Recovered from Patients in Egypt Compared to Global Strains. Microorganisms. 2023;11(8):1948.
49. Botelho J, Cazares A, Schulenburg H. The ESKAPE mobilome contributes to the spread of antimicrobial resistance and CRISPR-mediated conflict between mobile genetic elements. Nucleic Acids Res. 2023;51(1):236–52.
50. Tang Y, Fu P, Zhou Y, Xie Y, Jin J, Wang B, et al. Absence of the type I-E CRISPR-Cas system in Klebsiella pneumoniae clonal complex 258 is associated with dissemination of IncF epidemic resistance plasmids in this clonal complex. J Antimicrob Chemother. 2020;75(4):890–5.

Publisher's Note