Understanding the genetic diversity, antimicrobial resistance, and virulence of *Klebsiella pneumoniae* bacteria

Anton Spadar

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy

of the

University of London

December 2022

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

No funding received

Primary Supervisor: Professor Taane Clark

Secondary Supervisor: Professor Susana Campino

I, Anton Spadar, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Klebsiella pneumoniae (Kp) is among the top etiological agents of hospital acquired infections behind only *Escherichia coli* and *Pseudomonas aeruginosa*. Due to Kp's large accessory genome, it rapidly acquires antimicrobial resistance (AMR) genes in response to changing antibiotics used in clinical practice. Carbapenemase producing Kp are a particular problem as the resulting infections have very limited treatment options. Kp is usually described as a nosocomial pathogen, but there are hypervirulent strains of Kp (hvKp) endemic in East Asia, which cause community acquired pyogenic liver abscess. While usually susceptible to antibiotics, hvKp infections disseminate rapidly and require aggressive treatment.

In Kp, both AMR and hypervirulence are usually the consequence of horizontally acquired genes. Horizontal gene transfer is mediated by plasmids and might be inhibited by bacterial restriction-modification systems that have been suggested as a bacterial immune system. Restriction-modification systems generate methylation patterns that I have identified in clinical Kp isolates (n=8) using data from the PacBio third-generation sequencing platform. Long reads from this platform allowed me to create complete genome assemblies of these isolates.

I have also examined the evolving genomic patterns, including of AMR genes and plasmids, in a longitudinal collection of Kp isolates (n=509; years 1980 to 2019) sourced from Portugal that underwent short-read sequencing on an Illumina platform. The analysis revealed the active transmission of strains with AMR genes. A subsequent analysis of global Kp (n=725) characterised hypervirulence biomarkers in the core and accessory genomes using genome wide association study and machine learning methods. This analysis revealed not only known salmochelin and aerobactin loci, but also other genes putatively linked to hypervirulent phenotype.

Extending this work, I applied manifold learning and density-based clustering methods to all publicly available Kp assemblies (n=13,176) to investigate the relationship between carbapenemase genes, hypervirulence genes and plasmids. This analysis identified multiple likely outbreaks of carbapenem resistant hvKp and provided insights into the global dynamics of plasmids and genes they carry. In summary, my work has reinforced the importance of genomics and applied statistical methods to understand Kp hypervirulence, epidemiology, AMR and transmission.

Table of contents

Abstract 3

Table of contents 5

Acknowledgments 7

List of abbreviations 8

List of publications 11

Chapter 1: Introduction 13

Klebsiella pneumoniae 14

Klebsiella AMR drivers 16

Beta-lactams 16

Aminoglycosides 19

Fluoroquinolones 20

Other antimicrobial agents 21

Virulence factors 22

Capsule 22

Lipopolysaccharide 23

Fimbriae 24

Biofilm 25

Siderophores 25

Porins and efflux pumps 27

Mobile genetic elements 27

Insertion sequences and transposons 28

Gene cassettes and Integrons 28

Plasmids 29

Integrative conjugate elements 31

Methylation 32

Sequencing and Bioinformatic Analysis 33

Genome sequencing and data processing 33

Genome annotation 35

Population structure 35

Genotypic-phenotypic associations 39

Chapter 2: Research Questions 40

Datasets 43

Pacbio sequence data (n=8) 43

Illumina sequence data (n=509) 43

Publicly available data (n=13,176) 44

Chapter 3: Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals 46

Chapter 4: Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating antimicrobial resistance 59

Chapter 5: Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains 75

Chapter 6: Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains 86

Chapter 7: Discussion 115

Chapter 8: Conclusion 122

References 123

Appendix 1. Supplementary materials: Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals. 136

Appendix 2. Supplementary materials: Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating antimicrobial resistance. 148

Appendix 3. Supplementary materials: Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains. 155

Appendix 4. Supplementary materials: Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains. 163

Acknowledgments

I would like to thank my supervisors Prof. Taane Clark and Prof. Susana Campino. I would also like to acknowledge help and advice of my London colleagues Dr. Jody Phelan, Amy Ibrahim, Anna Turkiewicz, Emilia Manko, Holly Acford-Palmer, Dr. Leen Vanheer, Julian Libiseller-Egger, Sophie Moss, Dan Ward, and Gary Napier. I am also grateful for collaborative work with Portuguese colleagues Dr. João Perdigão, Rita Elias, Ana Modesto, José Melo-Cristino, Gabriela J. Silva, Constança Pomba, Cátia Marques, Margarida Pinto, Maria José Saavedra, Catia Caneiras and Prof. Aida Duarte.

List of abbreviations

AAC	aminoglycosides acetyltransferases
ABC	ATP-binding cassette
AME	aminoglycosides modifying enzymes
AMR	antimicrobial resistance/resistant
ANT	aminoglycoside nucleotidyltransferase
APH	aminoglycosides phosphotransferases
APH	phosphotransferases
CG	clonal groups
cKp	classic *Klebsiella pneumoniae*
CPS	capsule polysaccharide
CRhvKp	carbapenem resistant hypervirulent *Klebsiella pneumoniae*
Dam	DNA adenine methylase
DNA	deoxyribonucleic acid
ESBL	extended spectrum beta-lactamase
GC	guanine-cytosine
GI	gastrointestinal tract
GWAS	genome wide association study
HMV	hypermucoviscous
hvKp	hypervirulent *Klebsiella pneumoniae*
ICE	integrative and conjugative element

ICU	intensive care unit
InDel	insertion and/or deletion
IS	insertion sequence
Kbp	kilobase pairs
Kp	*Klebsiella pneumoniae*
KPC	*Klebsiella pneumoniae* carbapenemase
LPS	lipopolysaccharides
Mbp	megabase pairs
MDS	multidimensional scaling
MGE	mobile genetic element
MIC	minimum inhibitory concentration
MLST	multilocus sequence typing
Mtase	methyltransferase
NCBI	National Center for Biotechnology Information
NDM	New Delhi metallo-beta-lactamase
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PBP	penicillin-binding protein
PCA	principal component analysis
PCoA	principal coordinates analysis
R-M	restriction-modification
SNP	single nucleotide polymorphism

SNV	single nucleotide variant
ST	sequence type
TLR4	toll-like receptor 4
t-SNE	t-distributed stochastic neighbour embedding
UMAP	uniform manifold approximation and projection
UTI	urinary tract infection
WGS	whole genome sequencing

List of publications

Research papers in thesis:

[1] Spadar A, Perdigão J, Phelan J, Charleston J, Modesto A, Elias R, de Sessions PF, Hibberd ML, Campino S, Duarte A, Clark TG. Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals. *Sci Rep.* 2021;11(1):6491. doi: 10.1038/s41598-021-85724-2. PMID: 33753763.

[2] Spadar A, Phelan J, Elias R, Modesto A, Caneiras C, Marques C, Lito L, Pinto M, Cavaco-Silva P, Ferreira H, Pomba C, Da Silva GJ, Saavedra MJ, Melo-Cristino J, Duarte A, Campino S, Perdigão J, Clark TG. Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating antimicrobial resistance. *Sci Rep*. 2022;12(1):13791. doi: 10.1038/s41598-022-17996-1. PMID: 35963896.

[3] Spadar A, Perdigão J, Campino S, Clark TG. Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains. *Sci Rep.* 2022; 12(1):13671. doi: 10.1038/s41598-022-17995-2. PMID: 35953553.

[4] Spadar A, Perdigão J, Campino S, Clark TG. Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains. *Genome Medicine,* under revision.

Other papers accepted during PhD:

[5] Collins EL, Phelan JE, Hubner M, **Spadar A**, Campos M, Ward D, Acford-Palmer H, Gomes AR, Silva K, Ferrero Gomez L, Clark TG, Campino S. A next generation targeted amplicon sequencing method to screen for insecticide resistance mutations in *Aedes aegypti* populations reveals a rdl mutation in mosquitoes from Cabo Verde. *PLoS Negl Trop Dis.* 2022 Dec 13;16(12):e0010935. doi: 10.1371/journal.pntd.0010935. PMID: 36512510.

[6] Campos M, Phelan J, **Spadar A**, Collins E, Gonçalves A, Pelloquin B, Vaselli NM, Meiwald A, Clark E, Stica C, Orsborne J, Sylla M, Edi C, Camara D, Mohammed AR, Afrane YA, Kristan M, Walker T, Gomez LF, Messenger LA, Clark TG, Campino S. High-throughput barcoding method for the genetic surveillance of insecticide resistance and species identification in *Anopheles gambiae* complex malaria vectors. *Sci Rep.* 2022; 12(1):13893. doi: 10.1038/s41598-022-17822 8. PMID: 35974073.

[7] Elias R, **Spadar A**, Phelan J, Melo-Cristino J, Lito L, Pinto M, Gonçalves L, Campino S, Clark TG, Duarte A, Perdigão J. A phylogenomic approach for the analysis of colistin resistance-associated genes in *Klebsiella pneumoniae*, its mutational diversity and implications for phenotypic resistance. *Int J Antimicrob Agents*. 2022;59(6):106581. doi: 10.1016/j.ijantimicag.2022.106581. PMID: 35378228.

[8] **Spadar A**, Phelan JE, Benavente ED, Campos M, Gomez LF, Mohareb F, Clark TG, Campino S. Flavivirus integrations in *Aedes aegypti* are limited and highly conserved across samples from different geographic regions unlike integrations in *Aedes albopictus*. *Parasites & Vectors*. 2021; 14(1):332. doi: 10.1186/s13071-021-04828-w. PMID: 34174947.

[9] Perdigão J, Caneiras C, Elias R, Modesto A, **Spadar A**, Phelan J, Campino S, Clark TG, Costa E, Saavedra MJ, Duarte A. Genomic Epidemiology of Carbapenemase Producing *Klebsiella pneumoniae* Strains at a Northern Portuguese Hospital Enables the Detection of a Misidentified *Klebsiella variicola* KPC-3 Producing Strain. *Microorganism*s. 2020;8(12):1986. doi: 10.3390/microorganisms8121986. PMID: 33322205.

[10] Perdigão J, Modesto A, Pereira AL, Neto O, Matos V, Godinho A, Phelan J, Charleston J, **Spadar A**, de Sessions PF, Hibberd M, Campino S, Costa A, Fernandes F, Ferreira F, Correia AB, Gonçalves L, Clark TG, Duarte A. Whole-genome sequencing resolves a polyclonal outbreak by extended-spectrum beta-lactam and carbapenem-resistant *Klebsiella pneumoniae* in a Portuguese tertiary-care hospital. *Microb Genom*. 2019 Sep;7(6):000349. doi: 10.1099/mgen.0.000349. PMID: 32234124.

Chapter 1: Introduction

Klebsiella pneumoniae

Klebsiella pneumoniae (Kp) is a Gram-negative facultative anaerobic pathogen of genus *Klebsiella* in the *Enterobacteriaceae* family. Originally the sole member of the genus (1), the genus now includes 19 species (2). Kp is the most clinically relevant of these species, though *K. aerogenes*, *K. oxytoca*, and *K. variicola* are also encountered in clinical practice (3–5). The average Kp assembly is 5.6Mbp long and has 57.1% GC content (2). Of this, 5.4Mbp (ASM24018v2) is chromosomal and the remainder consists mostly of plasmids. Of the ~5,500 genes in a typical Kp genome, only ~2,000 constitute the core genome. The accessory genome of Kp is large with an apparent lack of convergence (6,7). The accessory genome makes Kp a pathogen of concern because it enables the bacterium to acquire antimicrobial resistance (AMR) and virulence genes that can be disseminated into other pathogens. One group of such genes encode carbapenemase enzymes that inactivate beta-lactam antibiotics. Notable among these is *Klebsiella pneumoniae* carbapenemase (KPC) first identified in 1996 (8) with multiple additional enzymes discovered thereafter including inhibitor resistant New Delhi metallo-beta-lactamase (NDM) in 2008 (9). The first documented KPC linked outbreak occurred in a hospital in the USA, with 14 patients infected and 8 deaths (10). Since it was characterised, KPC has spread rapidly throughout Americas, Europe, and Asia, while data availability is poor for Africa (11,12).

Kp is usually classified into strain-types based on multilocus sequences typing (MLST) inferred from seven core genes: *rpoB*, *gapA*, *mdh, pgi*, *phoE*, *infB*, and *tonB* (13). The notable sequence types (STs) include carbapenemase linked ST258 and ST11, as well as hypervirulent ST23 and ST86. Due to the growing availability of whole genome sequencing data, more detailed classification methods have been proposed. These are based on 694 or 1143 core genes have been proposed, but they have gained limited acceptance (14,15). Additionally, Kp strains are sometimes aggregated into clonal groups (CGs) based on isolates having fewer than 100 core genome SNPs between them, but the use of this scheme also remains limited (15). The difficulty of classifying the relationship between Kp isolates is its high propensity for recombination (16–18).

While commonly described as a nosocomial pathogen or commensal human bacteria, there is strong evidence that Kp is widespread in the environment. Based on 16S rRNA studies Kp is much more abundant in plants and insects (each 12% relative abundance) than in humans and agricultural animals (6 and 8% relative abundance, respectively) (19). Research outside of human health settings has shown that *Klebsiella spp* form the most abundant part of the metagenome of Mediterranean fruit flies (*Ceratitis capitata*), a pest of commercial orchards, and colonisation by *K. oxytoca* enhances mating competitiveness of the flies (20). However, beyond 16S rRNA studies there is little research on the environmental prevalence of Kp.

In humans, Kp commonly colonises the gastrointestinal tract (GI), but it is also frequently found in the nasopharynx. The prevalence of Kp colonisation varies substantially with estimates ranging from 4% GI and 10% nasal carriage in a community study in the USA (21) to 19% among Australian intensive care unit (ICU) patients who had a recent healthcare contact (22). Some studies report hospital patient colonisation rates of up to 38% (23). In Vietnam, the oropharyngeal carriage prevalence was 14%, but this varied substantially within the population. For example, the risk of Kp carriage is increased for smokers (odds ratio 1.9) and rural living (odds ratio 1.6) (24). Despite its name, Kp is not a major cause of community acquired pneumonia and accounts for <1% of cases requiring hospitalisation. Community acquired Kp pneumonia was more prevalent in the past, but the exact epidemiology is unclear (25,26). Urinary tract infection (UTI) is a more common manifestation of Kp, accounting for 9% to 17% of community acquired cases (27). A recent multiregional study found high variability of the primary infection sites among the cases of community acquired bacteraemia caused by Kp (26). Among 202 cases, 53 had pneumonia as the primary infection manifestation, but 92% of these came from South Africa and Taiwan. Liver abscess was identified as a primary manifestation in 18 cases, but 17 of these came from Taiwan. In contrast, UTI was the primary infection manifestation in 15% and 10% of community acquired bacteraemia in Taiwan and South Africa, respectively, but 38% (n=25) in cases in Europe, Australia, USA, and Argentina (26). Most of the nationwide monitoring is focused not on Kp itself, but on the AMR profiles of key bacterial pathogens including Kp (12,28,29).

Klebsiella AMR drivers

Kp is a key entry point for AMR genes (19). Its large accessory genome gives Kp access to a large number of AMR genes, and these acquired genes, rather than mutations, are the main drivers of AMR in Kp. Where data is available, the treatment of invasive Kp infections relies mostly on four antibiotic groups: fluoroquinolones, aminoglycosides, third generation cephalosporins, and carbapenems (**Table 1**). Carbapenemase inhibitors are also used if available (10,29).

Table 1. Selected antibiotics in current use and methods of administration (30–34)

Antibiotic	Introduced	Class	Administration	Resistance routes in Kp
Amoxicillin	1972	Penicillin	Oral	Mainly beta-lactamase enzymes (e.g., SHV, TEM, CTX-M, OXA), but also loss of porins
Ceftriaxone	1982	Cephalosporins	Intravenous, intramuscular	
Cephalexin	1969	Cephalosporins	Oral	
Imipenem	1985	Carbapenems	Intravenous	Carbapenemase enzymes (subset of beta-lactamases) e.g., KPC, NDM, VIM), but also loss of porins
Gentamicin	1964	Aminoglycosides	Intravenous	Mainly aminoglycosides modifying enzymes from three groups: AAC, ANT and APH
Levofloxacin	1996	Fluoroquinolones	Oral, intravenous, topical	Mutations in *gyrA* with likely compensatory mutation in *parC*
Colistin	1970 (agricultural use)	Polymyxin	Intravenous, intramuscular	*mcr* genes, mutations in *pmrB*, *crrB*

Beta-lactams

Cephalosporins and carbapenems are beta-lactam antibiotics that target cell wall biosynthesis by permanently blocking the active site of penicillin-binding proteins (PBPs). PBPs crosslink D-alanine-D-alanine which give rigidity to the peptidoglycan layer of the cell wall. This enzymatic activity is blocked by the beta-lactams, which permanently acetylate PBPs. The loss of enzymatic activity results in loss of cell wall rigidity (35–37). Of note, there are cell wall deficient forms of common pathogens including Kp, which are called L-forms. These L-forms are extremely resistant to beta-lactams even without normal resistance drivers (38).

Generally, resistance to beta-lactams is driven by either genes or, to lesser extent, by efflux pumps and loss of porins (33,35,37,39). All beta-lactamase enzymes act by hydrolysing the beta-lactam ring, thus inactivating beta-lactam antibiotics. Two beta-lactamase naming schemes exist: Ambler (based on structure) and Bush-Jacoby-Medeiros (based on specific function) (40,41). Here I use the Ambler classification which divides the beta-lactamases into Classes A, B, C and D. Classes A, C and D are further subdivided into broad-spectrum, extended-spectrum (ESBL) and carbapenemases. Class B has a very different structure from the others. It uses Zn^{2+} to hydrolyse the beta-lactam compared to serine in class A, C and D (33,36). Based on amino acid identity, the enzymes within each class are grouped into types. These types include SHV, TEM, KPC, CTX-M in Class A; VIM, NDM and IMP types in Class B; FOX, CMY and DHA types in Class C and OXA type in Class D. Alleles of the same enzyme are identified by a numeric suffix and can have different antibiotic specificity, as is the case with TEM-1 (Class A), TEM-3 (Class A ESBL), OXA-11 (Class D ESBL), and OXA-48 (Class D carbapenemase) (36).

Table 2. Beta-lactamase classes

Ambler class	**Other names**	**Examples of types**	**ESBLs**	**Carbapenemases**	**Examples of Inhibitors**
Class A	Penicillinase	CTX-M, TEM, SHV, KPC	CTX-M, TEM-10, SHV-5	KPC, GES-5	Avibactam, varobactam, relabactam for all, clavulanic acid except for KPC
Class B	Metallo-b-lactamase	IMP, NDM, VIM, SIM		All	None in clinical use
Class C	Cephalosporinase	FOX, DHA, CMY	All	None so far	Avibactam, varobactam
Class D	Oxacillinase	OXA	OXA-11, OXA-15	OXA-23, OXA-48; OXA-51, OXA-181, OXA-37	Avibactam for OXA-48-like, clavulanic acid otherwise

The classification of beta-lactamases into categories has important clinical implications due to different specificity of these enzymes to the same beta-lactam antibiotics. Beta-lactamases also vary in their susceptibility to beta-lactamase inhibitors - molecules that block beta-lactamases and allow antibiotics to work. Class A enzymes are the most common beta-lactamases with a diverse range of enzymatic activity. The most common TEM-type enzymes generally have no activity against third generation cephalosporins. The second most common enzyme type, SHV, can hydrolase third and fourth generation cephalosporins (42). Class A enzymes are efficiently and irreversibly inhibited by clavulanic acid which makes a penicillin plus inhibitor a potent therapeutic agent against class A carrying pathogens (43). The exception is KPC, which is the most common class A carbapenemase, and is globally widespread with two major variants (KPC-2, KPC-3). This enzyme type is inhibited by avibactam, but avibactam resistant clinical isolates have been reported (44).

Class B enzymes are all carbapenemases and are capable of hydrolysing penicillins, cephalosporins and carbapenems; although their activity against monobactams is limited. Class B enzymes are insensitive to clinically available inhibitors with only three compounds in clinical trials (43,45). Because of this, class B enzymes represent a serious threat, especially due to the growing incidence of NDM carrying *Enterobacteriaceae* including Kp.

Class C is chromosomally carried in *Enterobacteriaceae,* except in *Klebsiella* and *Salmonella* bacteria in which the encoding genes are carried by plasmids. These enzymes confer resistance to all beta-lactams except carbapenems. The enzymes are also not inhibited by clavulanate, but avibactam, varobactam and relabactam inhibitors are effective (43,46). Finally, Class D enzymes vary in their spectrum with OXA-48, OXA-23, and others capable of hydrolysing carbapenems. Class D enzymes are also very efficient inhibitors of oxacillin. These enzymes are only weakly inhibited by clavulanic acid; although OXA-48-like enzymes are inhibited by avibactam (36,43).

One further common route of beta-lactam resistance is through the loss or reduced expression of outer membrane porins such as OmpK35 and OmpK36 in Kp. Beta-lactams are generally unable to diffuse through the cell wall, and porins are the entry point (36,47). Mutations or recombination events that truncate porin genes lead to non-functioning proteins, and thereby increase the resistance in Kp. The mutations in genes regulating expression of porin genes (e.g., *kvrA*) likewise decrease susceptibility to beta-lactams via decreased uptake (48). However, given the universality of the porins their loss is likely associated with loss of fitness.

Because carbapenemases are the fastest growing group of resistance genes and carbapenemase producing Kp have very limited treatment options, my work focused on the epidemiology of this group (**Chapters 4**, **5** and **6**). North America and Europe are the regions with highest density of sampling, while most of Africa and South America only have single-centre studies.

In Europe, KPC and OXA-48 are predominant carbapenemases and they are endemic in Italy, Greece, and Turkey. NDM is spreading regionally in Poland and Romania with cases across Europe, and VIM is endemic in Greece (49) This dispersion presents an interesting insight into the geographic spread of carbapenemases. Despite freedom of movement within the European Union the absolute and relative prevalence of carbapenemases is highly variable.

Globally, KPC is the endemic carbapenemase in North and South Americas, and China, while NDM is endemic in Pakistan, China, Indian and Bangladesh. The endemicity of NDM in all four of these Asian countries is unexpected given much greater travel impediments compared to the European Union. The data for Africa is very limited, but suggestive of a regional presence of NDM in Kenya and Egypt, and endemicity of OXA carbapenemases in Morocco (47–49).

Aminoglycosides

Aminoglycosides are broad spectrum antibiotics that impede protein synthesis through binding to 16S rRNA of the 30S ribosomal subunit which in turn disrupts mRNA translation generating aberrant proteins (37). While originally very effective, the resistance to aminoglycosides is now widespread and is driven primarily by three types of aminoglycoside modifying enzymes (AMEs): nucleotidyltranferases (ANT), phosphotransferases (APH) and acetyltransferases (AAC), all of which modify aminoglycoside antibiotics. The nomenclature for these enzymes reflects the enzyme type, specific modification site and impacted antibiotics. For example, in AAC(6')-Ib, the "AAC" notes type of enzyme, "6'" the modification site, "I" is the specific resistance profile the enzyme confers, and "b" is an individual identifier (e.g., Ia, Ib). AAC is a catalyst in the acetylation of $-NH_2$ groups of the antibiotic. ANTs catalyse the transfer of an AMP group to aminoglycoside hydroxyl group. APHs catalyses the transfer of a phosphate group to the antibiotic (31).

The enzymatic activity of AMEs varies even within the same AME type, and their potency is antibiotic specific (53,54). A convenient example is the recently approved plazomicin, which is not affected by most AMEs except AAC(2')-Ia for which minimum inhibitory concentration (MIC) is 32 (µg/mL) compared to 2 (µg/mL) in bacteria lacking AAC(2')-Ia (54). The introduction of plazomicin also highlighted the resistance due to 16S rRNA methyltransferases encoded by *rmtB and armA* genes. Currently, these are relatively rare in Kp, but the use of plazomicin may drive their spread (55). There is an interesting relationship between different antibiotics, which is also observed in my investigation of Portuguese Kp samples **(Chapter 4)**. A large AMR phenotype dataset (n=278) of Kp and *Enterobacteriaceae* showed that >93% of ESBL producing Kp were susceptible to both plazomicin, amikacin and meropenem, but only 49% of these were susceptible to gentamicin. Among carbapenem-resistant *Enterobacteriaceae* 98% were susceptible to plazomicin, but susceptibility to amikacin and gentamicin has changed to 24% and 81%, respectively. As expected, only 3% of this cohort were susceptible to meropenem. Colistin-resistant *Enterobacteriaceae* had low amikacin, gentamicin and meropenem susceptibility rates of 21%, 12% and 12%, respectively, while for plazomicin it was 94% (56). A naïve expectation is that bacteria accumulate resistance genes, but here and in the Portuguese study (**Chapter 4**) this is clearly not the case with many isolates losing resistance to older antibiotics while acquiring resistance to newer ones.

Fluoroquinolones

Fluoroquinolones are inhibitors of the bacterial topoisomerases and thus inhibit bacterial DNA replication. To fit within the cell, bacterial DNA exists in a supercoiled form and access to DNA requires a change in the degree of coiling, which is mediated by topoisomerases I and II. The former reduces the number of negative supercoils while the latter adds them, thus unwinding DNA into a relaxed state, which is accessible to proteins (57). DNA gyrase and DNA topoisomerase IV are type II topoisomerase enzymes consisting of two copies of GyrA and GyrB or ParC and ParE, respectively. In Gram-negative bacteria, GyrA is the key target for fluoroquinolones, while in Gram-positive types it is ParC. Somewhat unusually for Kp, and perhaps reflecting the synthetic origin of fluoroquinolones, the resistance to these antibiotics is driven mainly by mutations in *gyrA* gene which encodes the key fluoroquinolone target in *E. coli* and likely in Kp (58,59). Resistance driving mutations have also been observed in *parC* but the resulting level of resistance is low. However, there are several other genes that result in some degree of resistance to fluoroquinolones. Carriage of efflux systems *oqxAB*, *qepA* and *qnr* alleles are all associated with resistance, as is carriage of aac(6')-Ib-cr (57). The latter gene is notable for also generating resistance to aminoglycosides – an example of uncommon cross-class resistance (60). Additional factors determining resistance are the loss of porins, which are required for absorption of antibiotics into the bacterial cell and upregulation of efflux pumps.

Other antimicrobial agents

Fosfomycin, tetracyclines, and colistin are additional agents sometimes used against Kp infections. Specifically, colistin was a preferred treatment agent for carbapenemase producing Kp until the widespread availability of beta-lactamase inhibitors (10). Fosfomycin is usually used for UTIs due to good oral bioavailability. There are some IV formulations, but their availability and use vary considerably between countries. In Kp and many other Gram-negative bacteria resistance is partly the result of core genome *fosA* gene, which encodes an enzyme inactivating the antibiotic. However, despite widespread use of fosfomycin over the past 40 years, the resistance rates remain below 10% for most regions suggesting a substantial fitness cost of resistance. Recently, plasmid carried *fosA* homologues have emerged which are much more efficient at inactivating fosfomycin (61–63). Colistin, also known as polymyxin E, and closely related polymyxin B were developed in 1950's but did not become widely used due to high nephro- and neuro-toxicity and variable pharmacokinetics. The emergence of bacterial strains resistant to all other major antibiotics made colistin a viable last resort treatment (10). As expected, wider use led to emergence of the resistance gene, *mcr*, which is carried by plasmids and has been disseminated globally (34,64). After beta-lactamase inhibitors for class A and C beta-lactamases became available, colistin use has fallen due to the inhibitor's better toxicity and pharmacokinetic properties. For class B and D beta-lactamase producing Kp, colistin remains a feasible last-resort option. One interesting aspect of colistin is that L-forms, which lack an outer membrane, are resistant to polymyxins just like they are intrinsically resistant to beta-lactams (65). Finally, tetracyclines are a family of antibiotics which, like aminoglycosides, target conserved 16S rRNA in the 30S ribosomal subunit. Except for carbapenemase producing isolates, tetracyclines are not generally used against Kp due to high levels of resistance. The exception is tigecycline which is sometimes used when other antimicrobials are unavailable (10,66–68).

Virulence factors

Kp has two key virulence phenotypes: classical (cKp) and hypervirulent (hvKp). Although specific numbers are unavailable, by far the most common phenotype is cKp, which is seen in hospital settings or in immunocompromised individuals. In contrast, clinical cases of hvKp usually start in the community in otherwise healthy individuals. The two phenotypes also vary in presentation. CKp is commonly linked to catheter, canula and mechanical ventilation and thus manifests as a UTI, bacteraemia, or pneumonia. HvKp usually presents as a pyogenic liver abscess with rapid metastasis to other organs (27,67,68).

Kp virulence factors generally belong to one of the five categories: capsule, lipopolysaccharide, fimbriae, siderophores and biofilms. For each of these categories there is a diversity of actual genes. For example, four different siderophore systems are frequently found in Kp. Enterobactin is present in >95% of Kp isolates, yersiniabactin is the next most common, while aerobactin and salmochelin are rare and generally found in hvKp isolates (27,70).

Capsule

The capsule is a polysaccharide layer that surrounds the bacteria and protects it from phagocytosis. Engineered strains lacking a capsule are much less virulent than wild types. The capsule consists of branched polysaccharides and the thickness of capsule varies between strains (27,72). The genes for production of capsule are in the *cps* operon, which encodes genes required for biosynthesis and export of capsule polysaccharides. The 5' and 3' ends of the operons are highly conserved and contain genes encoding proteins required for capsule polysaccharide (CPS) translocation and processing (*galF*, *cpsACP*, *wzi*, *wza*, *wzb*, *wzc*, *wbaP*/*wacJ*, *gnd*, and *ugd*). The middle of the operon is variable and encodes for proteins required for polymerization and assembly of CPS subunits. The common CPS typing system is based on the whole K-locus and identifies 77 different K-loci (73,74).

Some hvKp strains exhibit a hypermucoviscous (HMV) phenotype associated with capsule overproduction. Originally the HMV was thought to be the result of capsule overexpression, but subsequently found to frequently co-occur and not directly causal. The HVM phenotype requires production of regulator of mucoidy phenotype RmpA, which controls expression of *rmpC* that drives overexpression of capsule and *rmpD,* linked to HMV (69,72,75). Additional genes required for HMV are *kvrAB* and *rcsB,* all of which act as regulators of *rmpA* expression. Unlike *rmpACD*, both *kvrAB* and *rcsB* are part of the Kp core genome. In a potential link to the fitness cost of AMR, loss of *kvrA* leads to downregulation of porins OmpK35 and OmpK36, which reduces susceptibility to beta-lactams, but also downregulates capsule expression (48,75).

Lipopolysaccharide

Lipopolysaccharides (LPS) are another component of Kp's outer surface. They consist of three components: lipid A, core oligosaccharides and O-antigen polysaccharide. Lipid A anchors LPS to the outer membrane of bacteria and is highly conserved. Lipid A is a potent ligand for Toll-like receptor 4 (TLR4). Activation of this receptor leads to production of cytokines and chemokines, which recruit neutrophiles and macrophages to the site of infection. Modifications of lipid A may prevent its recognition by TLR4 and contribute to immune evasion and virulence. Capsule may also block recognition of lipid A by preventing access of TLR4 (71).

Like lipid A, the core oligosaccharides are conserved and serve as a link between lipid A and O-antigen. The latter is diverse and plays a role in the virulence of Kp by mediating immune evasion. The mutants lacking the O-antigen are more susceptible to complement mediated killing compared to wild types. Both activate the complement cascade, but in the wild type C3b binds to the O-antigen instead of cell membrane, thus prevents formation of pores which protects cell from lysis. (71,72)

Unlike the capsule, for which all related genes are part of same locus, LPS biosynthesis and export genes are in four different loci: *lpx* (lipid A biosynthesis), *waa* (core oligosaccharide biosynthesis), *rfb* (O-antigen biosynthesis) and *lpt* (export of LPS to exterior of the cell). The *rfb* cluster is the basis for classification of LPS into nine O-types (76). As with capsule types, the known O-types cover the majority of clinical isolates but may not reflect full diversity of the *rfb* loci in the environment (69,70).

Fimbriae

In Kp, type 1 and 3 fimbriae are associated with pathogenicity and are frequently found in clinical isolates. Both types contribute to biofilm formation but have different body site specificity. Type 1 fimbriae are thin strings with a thicker head. The strings, encoded by *fimA*, are attached to the Kp surface while head, encoded by *fimH*, can stick to D-mannosylated glycoproteins. Other proteins in the Kp type 1 fimbriae cluster encode for minor structural proteins (*fimF* and *fimG*), chaperon (*fimC*), usher protein (*fimD*), and uncharacterised proteins (*fimI* and *fimK*). Type 1 fimbriae are present in over 90% of clinical and environmental samples, but their expression is dependent on which part of the human body is colonised: the locus is expressed in UTI, but not in lungs or GI (71,77).

Type 3 fimbriae are helix-like filaments which are encoded by genes *mrkA* (main filament, binds to abiotic surfaces), *mrkBCE* (assembly and regulations) and *mrkD* (fimbriae head, binds extracellular matrices). This fimbria is also not required for colonisation of lungs or GI and its *in vitro* relevance to colonisation of the urinary tract is unclear. Both types help Kp bind to non-biological surfaces and extracellular matrices. The former allows Kp to form biofilms on medical equipment such as catheters and lung ventilation equipment. Binding to extracellular matrices allows binding to bronchial or urinary tracts (71,77).

Biofilm

Biofilms are bacterial communities bound by an extracellular polysaccharide matrix. Depending on Kp strains, the polysaccharide matrix contains elements of sugars from the capsule, O-antigen, or other sugars, which suggests large diversity of usable substrates. While type 1 and 3 fimbriae are both thought to contribute to biofilm formation, their main role is more likely to be anchoring of planktonic bacteria. The biofilm formation itself may not require attachment to fixed surface since floating biofilms, flocs, are observed in Kp (72,78,79). Kp biofilms likely increase resistance to phagocytosis and some antibiotics (77,78,80). Partly due to their diversity, the Kp pathways responsible for biofilm formation are poorly understood, with most focusing on the adhesion aspect of biofilm formation (72,81).

Siderophores

Siderophores are small molecules secreted by bacteria to scavenge for iron. In both the environment and humans soluble iron is scarce. In mammals, iron is usually bound to transport molecules such as haemoglobin or transferrin. The mammalian iron levels are further reduced in response to infection through binding of iron to lactoferrin. Compared to host iron binding proteins, the bacterial siderophores possess higher binding affinity for iron, which allows bacteria to scavenge iron from the host. The iron is imported into a cell by uptake of previously secreted siderophores. The siderophore systems normally consist of a biosynthesis locus, which encodes the siderophore molecule, the secretion and uptake proteins that reabsorb siderophores bound to iron (67,70,80,81).

Kp siderophores normally belong to one of four types: enterobactin (ent), yersiniabactin (ytb), aerobactin (iuc) and salmochelin (iro). Enterobactin is part of Kp core genome and among the four types has the highest affinity for iron. Biosynthesis and transport of enterobactin rely on *entABCDEF* and *fepABCDG* gene clusters, respectively, with *fepA* encoding uptake receptor. Enterobactin is inhibited by lipocalin-2, which helps clear the infection because bacteria cannot grow in the absence of iron. Because enterobactin is blocked by lipocalin-2, Kp strains that lack other siderophores cannot efficiently colonise lungs (84,85). This limitation is overcome by Kp strains carrying the second most common Kp siderophore, yersiniabactin, which is not inhibited by lipocalin-2 (84,86). The prevalence of yersiniabactin varies from 18% in cKp to 90% in hvKp. It is produced by proteins encoded by *irp* locus with transporters encoded by *ybt* and *fuy* and the uptake receptor encoded by *ybtQ*. While not susceptible to lipocalin-2, yersiniabactin has lower iron affinity than transferrin, which prevents iron acquisition by yersiniabactin carrying Kp (71). Both yersiniabactin and enterobactin are expressed in Kp during lung and respiratory tract colonisation, with the former allowing colonisation of lungs where lipocalin-2 inhibits enterobactin, and in turn enterobactin allows growth in blood plasma which contains transferrin (86,87).

Salmochelin is c-glycosylated form of enterobactin that evades lipocalin-2 (88). Salmochelin is relatively rare in cKp (2-4%) but is common in hvKp (>90%) (71). The reason for the rarity of salmochelin in cKp strains is unclear. Recent work that examined the growth of siderophore knockout mutants in human ascites and serum has determined that knockout of aerobactin, but not either one or all the other siderophores, prevents Kp growth. However, in a chicken infection model, aerobactin and salmochelin appear to compensate for each other's loss (87,88).

Aerobactin is infrequent in cKp (~6%), but is considered the key indicator of hvKp, as it is more frequent (>93%) in hvKp than salmochelin (71). Because aerobactin is often carried on the same plasmid as the genes required for HMV phenotype, the HMV and hvKp phenotypes were originally thought to be the same. However, the two are now considered as separate correlated phenotypes because the required genes are carried on the same plasmid (67,89).

An ABC iron transport system *kfu* is not a siderophore but is part of an iron acquisition pathway. This system, which consists of three genes, is reportedly linked to virulence in a mouse model. The mutant strain with *kfu* deletion did not cause mortality in the mouse peritonitis model (71,92). Using a representative global collection of Kp (n>12,000) (see **Chapter 6**), I found the *kfu* locus is perfectly correlated with ST (rho = 1), which makes it like type 1 fimbriae, but very different from the plasmid carried siderophores and resistance genes. This observation suggests there is limited selective pressure on the *kfu* locus.

Porins and efflux pumps

As the name implies, porin proteins form pores in bacterial membranes. In Kp, the porins OmpK35 and OmpK36 play role in both virulence and AMR. Loss of one or both porins leads to lower uptake of antibiotics into the cell and increases resistance to multiple agents including carbapenems. However, this enhanced AMR is balanced by lower virulence. The exact cause of lower virulence is unclear, but loss of porins may increase efficiency of phagocytosis, which leads to clearance of infection (32,43,69).

Efflux pumps export molecules from the cell, and in some cases, they can also expel antibiotics. In Kp, OqxAB and AcrAB are linked to AMR and the latter is also implicated in virulence. Mutants lacking *acrB* are attenuated in mouse models and have increased susceptibility to beta-lactams (71,93).

Mobile genetic elements

Mobile genetic elements (MGEs) may include both virulence and resistance factors. MGEs are a diverse group of genetic elements that allow transfer of genetic information both within and between bacterial cells. Phages are not normally classified as MGEs, but they facilitate transduction of MGEs by carrying them within a phage genome. MGEs can be generally split into those that are capable of independent movement between cells (plasmids and integrative conjugative elements (ICEs)) and those that can move between genomic regions of the cell but rely on other MGEs to move between cells (insertion sequences, transposons, integrons). The latter group creates diversity of plasmids and ICEs, as well as generating new combinations of genes in the same way random mutations result in greater fitness. Conjugation, exchange of DNA between cell via plasmids and ICEs, transduction, exchange of DNA via bacteriophages, and transformation, uptake of free extracellular DNA, are the main routes of horizontal gene transfer in bacteria (94). Both conjugation and transduction have been observed in Kp, with conjugative plasmids identified as major vector of AMR genes (19). While Kp may be capable of transformation, it is not known to be naturally competent (95). Due to the diversity of MGEs, only those that are relevant to Kp are covered here.

Insertion sequences and transposons

Insertion sequences (ISs) are small MGEs carrying one or two transposase (*tnp*) genes and range in size from 0.7 to 2.5 kbp. ISs can be grouped based on the amino acid residues of transposase at key cleavage site. The two types (DDE or HUH) cleave source DNA at 3' and 5' sites to generate single stranded DNA, which is then integrated into another genomic site. While the target of re-integration has some sequence specificity, the major target determinant may be the DNA conformation rather than nucleotide sequence. DDE transposases are the most common, while HUH types are part of the same family as plasmid relaxase proteins. When an IS carries a passenger gene, the unit consisting of IS and passenger genes is called a composite transposon (94,96–98).

Apart from moving AMR or virulence genes within cell, IS can directly affect genes by inserting into them which leads to a truncated reading frame or by affecting gene promoters (99). Gene truncation by IS is frequently observed in porin encoding OmpK35 and OmpK36 that contribute to reduced uptake of beta-lactams (100). The modification of gene expression due to an IS promoter has been observed in several AMR genes carried by Kp. Transposon Tn4401 contains ISKpn7 and ISKpn6. This transposon carries the globally dominant carbapenemase gene bla_{KPC}, which uses IS promoters (101). Similarly, carbapenemase encoding bla_{OXA-48} uses a promoter of the IS1999 (102) and carbapenemase bla_{NDM-1} uses an IS*Aba125* promoter (103). Another IS, *ISApl1*, transposes and supplies the promoter for colistin resistance, *mcr-1*, into many bacterial species including *E. coli* and Kp (34).

Gene cassettes and Integrons

Gene cassettes are small MGEs (0.5 to 1kbp) that usually cannot self-replicate. The cassettes normally contain a recombination site, *attC*, and a gene that lacks a promoter. Cassettes are usually located within integrons – an MGE consisting of an integrase (*intI*) gene and *attI* recombination site. Integrase catalyses recombination between the cassette's *attC* and integron's *attI* sites, which can lead to an integron acquiring one or more cassettes. Integrons also contain two promoters: one to drive expression of *intI* and one to drive expression of the cassette genes, which are normally inserted in the opposite orientation to the integrase gene (94,104).

Gene cassettes carry a wide range of genes including AMEs, carbapenemases *bla_{IMP}* and *bla_{VIM}*, but to date there is no comprehensive database listing these. Neither the Repository of Antibiotic resistance Cassettes nor Multiple Antibiotic Resistance Annotator, both of which list gene cassettes, are currently accessible (95,102).

Plasmids

ISs, transposons, gene cassettes and integrons move genes between sections of DNA within bacteria, but generally not between cells. However, they can embed themselves within MGEs that can move between cells. This movement is usually performed by plasmids, ICEs, or bacteriophages. While movement within species is most prolific, between species movement of plasmids is also widespread (106). Plasmids are large extrachromosomal genetic elements which are capable of self-replication. Like bacteria which carry them, the plasmids have a core genome that encodes for the key functions of replication and conjugation, though only a quarter of plasmids appear capable of self-conjugation (107). In another similarity to other bacteria, plasmid replication starts at the origin of replication. While plasmids carry a gene encoding for replication initiation protein, they usually rely on a host's helicase, polymerase, and other replication machinery. This may explain the observed host specificity of the plasmids. The horizontal transfer by conjugation relies on plasmid encoded type IV secretion system, which assembles a pilus that connects to a recipient cell; though not all plasmids carry the genes required for conjugation (94,107–109).

Unlike the host's chromosome which exists as a single copy, a bacterial cell carry multiple copies of the same plasmid. This increases the opportunity for those MGEs which cannot self-replicate to replicate and relocate between DNA regions.

Because they facilitate between cell transfer of resistance genes, plasmids are the focus of the attempts to classify MGE driven AMR. The original plasmid classification scheme relied on the observed incompatibility of closely related plasmids. Copy control systems of closely related plasmids can mistakenly act on non-self plasmids and limit plasmid replication before the optimal number of copies per cell is reached. This in turn causes inefficient segregation between dividing cells. The incompatible plasmids were assigned to the same Inc groups, but incompatibility testing is impractical at a large-scale, so the current typing methods rely on *in-silico* classification. One popular method, PlasmidFinder, uses plasmid replicons (origin of replication *ori* and *rep* gene encoding replication initiation protein) while another, MOB-suite, adds relaxase proteins to the replicon (110–112). Confusingly, a replicon-based scheme has retained the old Inc-based naming convention. The classification is made more difficult by some plasmids carrying more than one replicon or mobility locus.

Kp plasmids carrying resistance and virulence genes are quite diverse and form part of large *Enterobacteriaceae* plasmid ecosystem (106). Two recent reviews of plasmids carrying AMR genes identified over 55 Kp specific plasmids of which 37 carried beta-lactamases. Based on replicon typing these plasmids belonged to all common replicon groups (113). Similarly, a review of AMR plasmids in *Enterobacteriaceae* found many plasmids carrying carbapenemase genes in all major plasmid replicon groups, including 98 in IncF, 72 in IncI, 9 in IncA/C, 16 in IncH and 240 among other plasmid types (114). The problem of enumerating the Kp AMR plasmids is apparent even with this short list: several of the plasmids are a result of recombination as supported by BLAST comparison results **(Table 3)** (115). Two FIIk replicon plasmids pKpQIL-234 and pKpQIL have 99% identity over the length of plasmids, although a ~23 kbp segment is inverted in one relative to the other. Each of these carries one of the two most common alleles (*bla*$_{KPC-2}$ and *bla*$_{KPC-3}$) of the carbapenemase encoding gene. Both plasmids also have 98% identity and 41% coverage against pKP09085, which carries only ESBL encoding *bla*$_{CTX-M-15}$. Similarly, two X3 replicon plasmids (pJEG027 and pKpS90) have >99% identity across >85% of their length, but one carries a carbapenemase gene *bla*$_{NDM-1}$ and the other *bla*$_{KPC-2}$, which encode structurally very different carbapenemases. Finally, N1 plasmids pNL194 and Plasmid 9 have >98% identity in >54%

of their length, but one carries *bla*$_{VIM\text{-}1}$ and the other *bla*$_{KPC\text{-}2}$, which are again metallo-beta-lactamase and class A carbapenemase encoding genes. Plasmids are also carriers of important Kp virulence genes with hvKp usually carrying *iro*, *iuc*, and *rmp* loci on plasmids such as pLVPK and Kp52.145 (116–118)

Table 3. Epidemic Kp plasmids carrying beta-lactamases. (113)

Plasmid reference	Accession no.	Resistance gene/s	Replicons	Countries of isolation
pKpQIL-234	KJ146689	*bla*$_{\text{KPC-2}}$	FIIk, FIB$_{\text{pKpQIL}}$	USA, Greece, China
pNDM-KN	JN157804	*bla*$_{\text{NDM-1}}$	A/C2 (R)	USA, Canada, Australia, Kenia, Taiwan
pJEG027	KM400601	*bla*$_{\text{NDM-1}}$	X3	Canada, China, India, USA
Plasmid 9	FJ223607	*bla*$_{\text{KPC-2}}$	N1 (A/C2, R)	Brazil, USA, China, Russia, Taiwan
pKpn-E1.Nr7	KM406491	*bla*$_{\text{OXA-48}}$	L	Switzerland, Ireland, Australia, France, Netherlands
pNL194	GU585907	*bla*$_{\text{VIM-1}}$	N1 (R)	Greece, Norway, Switzerland
pKP09085	KF719970	*bla*$_{\text{CTX-M-15}}$	FIIK, FIB$_{\text{pKpN3}}$	China, Sweden, Korea, Norway
pKpS90	JX461340	*bla*$_{\text{KPC-2}}$	X3 (U)	Brazil, Hong Kong, France, Italy
pKpQIL	GU595196	*bla*$_{\text{KPC-3}}$	FIIk, FIB$_{\text{pKpQIL}}$	USA, Israel, Italy, Taiwan

Integrative conjugate elements

Unlike plasmids that exist separately from host bacteria DNA, ICEs are integrated in the host genome and vary in size from 18 to 500kbp. Like plasmids, ICEs rely on type IV secretion system to enable movement between cells. The ICE conjugation genes are normally expressed at very low levels or not at all. When the change in environmental conditions induces the expression, the encoded proteins excise ICE from the host chromosome, which allows ICE to reintegrate into a different region of the host's DNA or conjugate to a different host via a type IV secretion system. The integration usually takes place at an *attB* site in the host chromosome and is catalysed by an ICE carried integrase. However, due to similarity of plasmids and ICE's, some of the catalytic and conjugation machinery is interchangeable (94,119).

ICEKp and its structural variants are the main ICEs in Kp. They all carry the *ybt* locus that encodes biosynthesis of the yersiniabactin siderophore. Additionally, some structural variants of ICEKp also carry a *clb*/*pks* locus encoding colibactin, a genotoxic polyketide (120).

Methylation

Methylation is a form of epigenetic modification which is ubiquitous in both prokaryotes and eukaryotes. The formation of epigenetic lineages enables the adaptation of bacterial populations to changing environments and modulates the interaction of pathogens with their eukaryotic hosts (121–123). Epigenetic signals control DNA–protein interactions and can cause phenotypic change in the absence of mutation (124). A common mechanism of epigenetic signalling is DNA methylation by orphan methyltransferases (MTases) such as DNA adenine methylase (Dam), which has roles in chromosome replication and segregation, nucleoid organization, cell cycle control, and DNA repair (124–126). DNA methylation is also the key element of Restriction-Modification (R-M) systems that not only provide a defence against foreign DNA, but also drive bacterial evolution by ensuring persistence of plasmids and other MGEs (127,128).

Methylation cannot be observed in short-read sequencing data, but if PCR is not used for library preparation, some long-read sequencing technologies allow detection of DNA methylation signatures by measuring variation in either the polymerase kinetics (Pacific Biosciences) or voltage-bias during sequencing (Oxford Nanopore Technology) (129). Here, I focus on Pacific Biosciences (PacBio) instruments (**Chapter 3**). This approach can detect genome-wide MTase N6-methyladenine (m6A) and N4-methylcytosine (m4C) target motifs at coverage levels recommended for assembly and reveal phase variation of related genes (130). Dam, which methylates m6A in the GATC sequence, plays a key role in DNA mismatch repair and in bacterial virulence and gene expression including in some strains of Kp (127,131–133). R-M systems have also been observed on bacterial plasmids where they may contribute to plasmid maintenance (128). Kp is known to have type I and type II R-M systems (134). The two systems have different mechanics and their distinct motif types have been reviewed elsewhere (124,126). Briefly, type I R-M systems consist of specificity, modification, and restriction subunits. The first two subunits are usually located together on a chromosome and are under control of same promoter (135). The restriction subunit in the type I R-M system recognises long bipartite motifs, such as GGCAN$_8$TCG. While not part of complete type II R-M, Dam is the most common MTase in Gamma-proteobacteria and it recognises palindromic 5'-GATC-3' motifs (124,126).

Sequencing and Bioinformatic Analysis

As next generation sequencing technologies (e.g., Illumina platforms) have become widely available they have improved the completion of reference genomes. This gave insights into genome diversity leading to the tracking of infections, development of diagnostics, and insights into vaccine design (e.g., for SARS-CoV-2 (136)). In Kp, whole genome sequencing (WGS) has led to an understanding of the bacteria's accessory genome (70) and elucidated the diversity of AMR drivers. This knowledge has enabled the development of tools to rapidly screen isolates for these drivers (137,138). However, next generation sequencing platforms usually cannot separate different DNA molecules (chromosome, plasmids, transposons) due to short read lengths and consequently limit understanding of the diversity of plasmids that transfer AMR genes between bacterial cells. The advent of third-generation sequencing via technologies from PacBio and Oxford Nanopore Technology (ONT) fills this gap, because they provide much longer (>10,000nt) sequencing reads. These devices also give insights into bacterial methylomes as they capture methylation data during sequencing (122,139).

Genome sequencing and data processing

Currently WGS is performed via either short or long-read sequencing. The former is primarily carried out using Illumina platforms. Short-read sequencing has error rates below 0.1% per base and allows the accurate calling of variants with low sequencing coverage of target regions (140). However, the read length is limited to a current maximum of 600nt, which means this type of sequencing cannot resolve long repetitive regions of DNA. Long-read sequencing is performed using either PacBio or ONT platforms. As the name implies, the main advantage of long-read sequencing is the ability to produce reads over 10,000nt in length. However, this comes at the cost of accuracy, with reported error rates for PacBio and ONT machines of 0.5% and 7%, respectively, per base (140,141). An extra advantage of both long-read sequencing platforms is that using the information collected during sequencing PacBio and ONT proprietary software allows detection of DNA methylation. In case of PacBio, methylation detection is part of their SMRT Analysis toolkit, which also includes quality control and other pipeline tools. The toolkit can be computationally intensive and slow for many samples, so for my thesis, I have adapted the key code file of the toolkit (*KineticWorker.py*) to improve analysis automation.

Long and short sequencing reads are processed in two main ways: alignment or assembly. For alignment the sequencing reads are mapped to a representative genome of the species and the differences between sequencing reads and reference genome analysed. Several multiple aligners have been developed for short reads (e.g., BWA, Bowtie2, Hisat2, and STAR software), most performing comparably well (142,143). Using the alignments, single nucleotide polymorphisms/variants (SNPs/SNVs), insertions and deletions (indels) and other variations can be called using established software tools (e.g., GATK, Samtools/BCFtools, or VarDict (144–146)). This mapping approach is common for stable genomes such as human, mice or *M. tuberculosis.* While an alignment approach can be used for Kp, due to high diversity of its genome, I opted to perform reference-free assembly of the genomes of each isolate. Genome assembly refers to the process of taking DNA sequencing reads and putting them back together to create a representation of the original chromosomes from which the DNA originated. In this thesis (see **Chapters 4** and **5**), short Illumina reads were assembled using Unicycler software (v0.4.8) (73). The subsequent analysis relies on comparison of regions (e.g., genes, promoters) of these assembled genomes. Because I used reference-free assembly in the analysis (see **Chapters 3, 4** and **5**), SNP detection was performed using alignments of the assembled genomic regions (e.g., using MAFFT, MUSCLE software).

Assembly of long reads (from PacBio sequencing; see **Chapter 3**) was performed using three different software tools: HGAP3, Canu and Flye. HGAP3 is part of the SMRT Analysis toolkit (v2.3), while Canu and Flye software are standalone (147–149). Typically, short read data can be used to polish the assemblies, but as each Kp isolate sequenced had high (>100-fold) long read coverage, this was not required. To measure the quality of assemblies from each software tool, N50 scores (related to the median and mean length of a set of sequences), and Busco gene completeness and fragmentation can be estimated. For Kp, I estimated the latter using the Busco Enterobacterales (odb10) dataset (150) to find assemblies with the highest sum of complete and fragmented genes, which also had high N50, with most arising from a pipeline involving Flye assembly. For the short read assemblies, the fragmentation and completeness were assessed against 440 core genes of Enterobacterales (enterobacterales_odb9 for **Chapter 4**; enterobacterales_odb10 for **Chapter 5**) using Busco software (v4) (150). The quality of assemblies was high, with all but one having complete single copies of >95% genes in the Busco reference set.

Genome annotation

Gene annotation needs to be inferred for an assembled genome and is typically transferred from a similar genome. Here, the assemblies were annotated using Prokka software (v 1.14.6) (151), combined with the *Klebsiella* specific reference genes set (152) (see **Chapters 3, 4** and **5**). O and K antigen serotypes*,* genomic AMR, virulence, and ST profiles were analysed and inferred *in silico* using Kleborate (118) and AMRFinder (v3.8.4) (137) software with associated databases. I used Abricate software (v1.0.1) (153) with the virulence factor database vfdb (accessed March 2021) to find additional virulence genes. The plasmid identities were established using PlasmidFinder (110) and ISs were identified using ISFinder (154).

Population structure

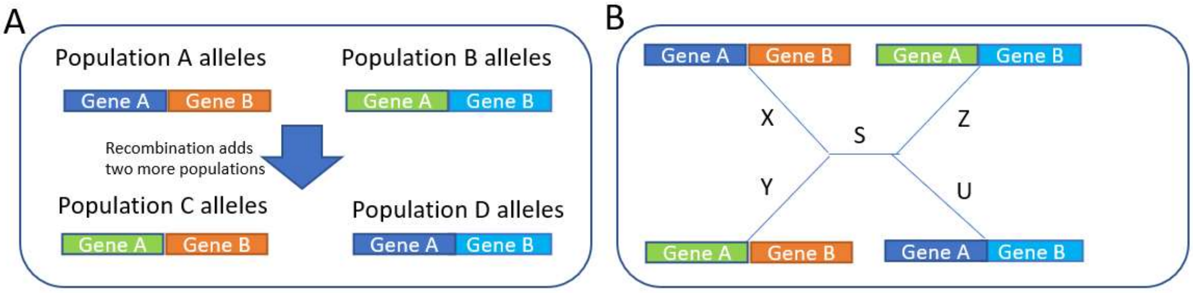
Understanding the relationships between isolates in populations can reveal sub-populations/clades linked to transmission, AMR, virulence, and strain-types. For example, isolates with (near-)identical genomes, sourced from different individuals, may reflect a transmission cluster or outbreak. The most common way to establish relationship between bacterial isolates is via phylogenetic trees. The approach usually follows one of two routes. The first route is to align sequencing reads to reference genome to identify SNPs (see above), which are inputted into phylogenetic reconstruction tools (e.g., RAxML and IQ-TREE2 (155,156)). As a part of the tree reconstruction process, these tools also provide bootstrap estimates as a measure of confidence that tree nodes are robust. The second route starts with identification of genes (or proteins) on which the phylogenetic tree will be based. The sequences from all isolates are aligned against each other using multiple sequence alignment tools such as MAFFT, MUSCLE or (for proteins) ProbCons (157–159). This multiple sequence alignment is again inputted to either RAxML or IQ-TREE2 (155,156). The advantage of first method is that isolates do not need to be assembled, while the second route gives better control over the data underlying the phylogenetic tree. For example, excluding highly variable or repetitive regions is easier via the second route. Phylogenetic tree reconstruction in this thesis applies the second route, using a subset of core genes (27 genes in **Chapter 3,** 100 genes in **Chapter 4**) with the most diverse nucleotide sequences. For each of these genes, I aligned the instances from all isolates using MAFFT software (v7.467) and a phylogenetic tree was reconstructed using IQ-TREE (v2.0.3) (156,157).

The application of phylogenetic trees in bacterial genomic has a long history, including for inferring evolutionary relationships among distinct taxa though not without a set of limitations (160,161). Phylogenetic trees are also widely used in viral genomics where recombination is a lesser concern (162), but in bacterial genomics their accuracy has been questioned (18,163). One of the major limitations of phylogenetic trees is their inability to reflect sexual reproduction or horizontal gene transfer (18,163). Several solutions have been proposed: removing or adjusting the trees for recombinant parts (164,165), creating graphs for subgroups of samples (166), modifying the definition of the phylogenetic tree to allow visualisation of recombination (167,168), and relying on a core genome (163,169). The non-robustness of a phylogenetic tree in the presence of recombination can be demonstrated using a simple example. Assume, there are two distinct populations of an organism, each with unique versions of two genes A and B. At some point, the two populations come into contact and recombination creates two more genotypes **(Figure 1A).** The observer who sees all four populations and tries to build a phylogenetic tree will not succeed as no tree can accurately capture the relationship between isolates. **Figure 1B** shows a generic unrooted phylogenetic tree with branch lengths X, Y, S, Z and U. The distance between all leaves of this tree can be summarised by the following equations where the right-hand side is the number of allele differences between samples at either end of the path:

$$\begin{gathered} X + S + Z = 2 \\ X + S + U = 1 \\ X + Y = 1 \\ Y + S + Z = 1 \\ Y + S + U = 2 \\ Z + U = 1 \end{gathered}$$

This system of equations is inconsistent i.e., there are no S, U, X, Y, and Z values for which all equations are satisfied. Therefore, the distance between leaves cannot be correct.

Figure 1. Example of simple recombining population (A) and generic phylogenetic tree in which distances between isolates will be incorrect regardless of the lengths [S, U, X, Y, Z] of the individual the branches (B)



Alternative approaches that look at population structure involve dimensional reduction, such as principal component analysis (PCA). PCA converts a data matrix into a covariance matrix, calculates eigenvectors and creates new “orthogonal” dimensions from it. The original data is then projected onto these new dimensions of which the first two (PC1 and PC2) are usually displayed. A common extension of PCA is multidimensional scaling (also known as MDS, principal coordinate analysis or PCoA) in which covariance matrix is replaced by distances between isolates often obtained from phylogenetic trees. MDS can accommodate ordinal data making it useful for analysis of categorical data such as multiple sequence alignments, but the calculations become more convoluted and the result dependent of choice of optimisation function. In addition, like PCA, MDS is not designed to deal with non-linear data structure (170,171).

A prominent group of dimensional reduction algorithms is based on manifold learning – an approach that assumes that complex data lies of smooth low-dimensional surface. The goal of manifold learning algorithms is to faithfully project that surface into lower dimensional (usually 2) space. Examples of manifold learning methods include ISOmap, t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP) (172–174). These approaches are usually better at representing the data structure than commonly used PCA, but they are not without limitations. The main analytical drawback is the need to select a set of parameters which vary between algorithms. For example, a wrong choice of distance metric may force unrelated points together giving impression of relatedness. Another drawback is possible distortion of embedded data (172,173,175). Manifold learning methods assume that data lies on low-dimensional surface, but there is no theoretical guarantee that this surface can be faithfully represented in two dimensions. A useful application of manifold learning methods is to understand if some structure exists within data, and subsequently identify the variables that drive that structure. Despite these drawbacks, manifold learning technics, especially UMAP and t-SNE algorithms, provide better understanding of data structure compared to PCA, MDS and some other methods (175–177).

In my research I have made extensive use of t-SNE and UMAP as they perform well and have acceptable computational time. Just like other manifold learning methods t-SNE and UMAP require a distance metric. The most frequently used, *Euclidean*, given by $d = \sqrt{\sum_{i=0}^{N}(x_i - y_i)^2}$ is unsuitable for categorical or Boolean data. For categorical data, such as those from multiple sequence alignments, the appropriate metric is *Hamming* defined as $d = \sum_{i=0}^{N}(1\ if\ x_i! = y_i\ else\ 0)/N$. However, in case of Boolean data (where 0 is absence of feature and 1 is presence) this metric may cause samples which have no similarity to anything to be grouped together. A better choice is often *Jaccard* or *Russell-Rao* defined respectively as

$$d = \sum_{i=0}^{N}[1\ if\ (x_i + y_i = 1)\ else\ 0] \bigg/ \langle N - \sum_{i=0}^{N}[1\ if\ (x_i + y_i = 0)\ else\ 0]\ \rangle$$

and

$$d = \langle N - \sum_{i=0}^{N} [1 \; if \; (x_i + y_i = 2) \; else \; 0] \, \rangle \Big/ N.$$

The downside of the *Jaccard* metric is that total number of possible features is disregarded because the denominator is the number of features present between the two samples. In contrast, the denominator of the *Russell-Rao* metric is the number of possible features, but this may create large distances in matrices with large number of possible features, but few features per sample.

For population structure analysis (**Chapter 4, 5 and 6)**, a matrix of the presence/absence of SNPs, accessory genes or plasmid replicons was analysed using UMAP (v0.5.1) and t-SNE algorithms using *Russell-Rao*, *Jaccard* and *Hamming* distances. This approach was implemented in sklearn software (v0.24.2) (174), and used to generate a 2-dimensional projection of the original binary matrices (172–174). HDBSCAN software (v0.81) was used to detect clusters in the 2-dimensional projections (178).

Genotypic-phenotypic associations

Genome wide associations studies (GWASs) are a method of identifying genomic features that are associated with an organism's phenotype. In bacteria, GWAS is commonly used for identification of AMR genes and mutations (179). In typical GWAS, for each genomic feature (e.g., SNP, gene) one calculates the feature's measure of non-random association with phenotype. GWAS approaches can range from simple statics like chi-squared analysis of two-way tables, to complex models involving non-genomic variables, surrogates of isolate ancestry (e.g., kinship matrices or principal components), and other features or confounding variables. Commonly used GWAS software packages are Plink, pyseer and treeWAS (180,181). In **Chapter 5**, I screen accessory and core genome SNPs for association with a hypervirulent phenotype using a logistic regression implemented in statsmodel software (v.0.13.0). The model incorporated principal components as variables that summarise population clusters, where the PCs were inferred from a PCA applied to core genome SNPs (182). The logistic model leads to epidemiological interpretable odds ratios of the effects of alleles after correction for confounding effects (e.g., population structure).

Chapter 2: Research Questions

The overarching goal of the research contained in this thesis is to understand the parts of Kp accessory genome related to AMR, virulence, methylation, and plasmids. For this purpose, I have examined the methylome and AMR profile of eight Kp isolates from Portugal (**Chapter 3**). I then conducted a large study of Portuguese Kp isolates (n=509) that gives insight into AMR and plasmids present in Portugal during the 40-year period covered by the isolates **(Chapter 4**). After that study, I have completed an investigation of biomarkers linked to hypervirulence in Kp using both core genome SNPs and accessory genome **(Chapter 5)**. I brought these three studies together through a fourth study **(Chapter 6)** that gives insight into the transmission of AMR and hypervirulence genes by plasmids in global context **(Table 4)**. The summary of methods used in each Chapter (research paper) are provided **(Table 5)**, and described in **Chapter 1**.

Table 4. Research papers included in this thesis.

Chapter	Manuscript title	Topic	Citation
3	Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals	Characterising the methylome of Kp (n=8)	Spadar et. al, *Sci Rep* 2021; 11(1):6491
4	Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating AMR	Understanding the local AMR and virulence genes transmission through whole genome diversity in Portugal (n=509)	Spadar et. al, Sci Rep 2022; 12(1):13791
5	Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains	Determine biomarkers of Kp hypervirulence (n=725)	Spadar et. al, *Sci Rep* 2022; 12(1):13671.
6	Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains.	Examine relationship between plasmid replicons, AMR and hypervirulence genes (n=13,176)	*Genome Medicine,* revised (2022).

Table 5. Summary of methods used

Short title	**The Kp methylome**	**Genomic investigation of Kp strains resistant to antimicrobials in Portugal**	**Genetic barcodes of hypervirulent *Kp***	**Large scale genomic analysis of global Kp plasmids**
Chapter (Appendix)	3 (1)	4 (2)	5 (3)	6 (4)
Focus	Methylation	AMR genes, temporal changes	Hypervirulence biomarkers	Plasmids, AMR and hypervirulence genes
Sample Sequencing	Yes	Yes	No	No
No. of Samples	8	509	N/A	N/A
Public Data	Yes	No	Yes	Yes
No. of Samples	83	N/A	725	13,176
Total Samples	91	509	725	13,176
Assembly	HGAP/Canu/Flye	Unicycler	Unicycler	N/A
Quality Assessment	Busco DB v10	Busco DB v9	Busco DB v10	Kleborate
Phylogenetic trees	IQ-Tree (26 genes)	IQ-Tree (100 most diverse core genes)	N/A	N/A
MLST profiling	Kleborate	Kleborate	Kleborate	Kleborate
AMR Profiling	Kleborate	Kleborate, AMRFinder	Kleborate	Kleborate
MGE Profiling	PlasmidFinder, ISFinder	PlasmidFinder	PlasmidFinder	PlasmidFinder
Virulence Profiling	N/A	Kleborate, Abricate	Kleborate	Kleborate
Annotation	Prokka	Prokka	Prokka	N/A
Core genome definition	MLST loci + 19 other	+/- 1% median length and pairwise identity >99%	+/- 20% median length and >90% pairwise identity and present in >90% of samples	N/A
Homologous gene definition	N/A	N/A	+/- 20% median length and >60% pairwise identity	N/A

Datasets

Pacbio sequence data (n=8)

For **Chapter 3**, I analysed Pacbio sequence data for eight Kp isolates. These eight samples were collected from Hospital de Santa Maria (n=6) and Hospital St Antonio dos Capuchos (n=2) in Lisbon, Portugal. The samples were cultured and tested for AMR as described previously (183). DNA was extracted from strain cultures, grown overnight at 37ºC on Mueller-Hinton Agar. DNA extraction was carried out using the Cetyl trimethylammonium bromide method previously described using Tris-Acetate buffer (10 mM, pH 8) (184). The samples were generated without TET1 oxidation; thus, I could not examine 5mC methylation (130,185). All DNA samples were sequenced at the Genome Institute Singapore on the PacBio RSII platform.

The eight samples in this dataset were selected to represent a range of dates (years 2005-2013) and different STs. While this gave a diverse set of restriction-modification enzymes, the small size of the dataset limits the power of the analysis. The diversity of STs is also a limitation because it increases heterogeneity between isolates further limiting power of small dataset. I have examined public databases for data that can be used to extend my analysis and identified a suitable dataset (PRJEB6403) containing 163 isolates, but it proved too difficult to establish who owned this unpublished data and establish a collaborative project.

Illumina sequence data (n=509)

For **Chapter 4**, Illumina data was generated for 509 Kp isolates. These isolates were identified between years 1980 and 2019 from 16 hospitals in Lisbon and its metropolitan area (Southern region), Coimbra (Central region), and Porto and Vila Real (Northern Portugal), except for 9 isolates from Beirolas wastewater (Lisbon) and 41 samples from veterinary clinics (Lisbon)**.** Biochemical identification methods (e.g., API, Vitek) were used. DNA was extracted from strain cultures, and whole genome sequencing was performed on Illumina HiSeq (paired end 150bp) platform. AMR testing data was available for all but 388 isolates.

The isolates in this dataset were selected to be a good representation of Portuguese clinical isolates. However, the period from 1980 to 1999 has limited coverage with only 13 isolates from 1990 to 1999. While this limits the historical value of the dataset, the focus on the period from 2000 to 2019 gives better insight into AMR likely to be facing clinicians. The dataset also has a skew towards isolates of greater interest to clinicians, which manifests as overrepresentation of KPC-3 encoding samples (20%).

Publicly available data (n=13,176)

The work in **Chapters 3, 5** and **6** used public data downloaded from the NCBI database (115,186). For **Chapter 3**, I have compared the newly sequenced isolates to all NCBI RefSeq database isolates (n=3,584) with a subset of these (n=83) used for phylogenetic reconstruction. In **Chapter 5**, to identify isolates sourced from liver and therefore likely to be hypervirulent, I have used the keywords “live” and “hepa” to search the metadata of samples in the NCBI Isolates Browser [n=79] (187). I supplemented these with 520 randomly chosen isolates from the same database and a geographically concentrated set of isolates (n=126) from three hospitals in Thailand. For **Chapter 6**, I have used all isolate assemblies (n=13,176) labelled as Kp that were available at the time in NCBI Isolates Browser (Sep-2021).

The isolates from public databases used in **Chapter 3** and **Chapter 6** represent all Kp isolates that were available respectively in NCBI RefSeq and Isolates Browser databases at the time of analysis. This data has biases inherent in such databases: dominance of small set of countries and mostly clinical samples. However, the advantage of using all public data is a dataset with improved statistical power and wider context. The data in **Chapter 5** is different because prior to randomly selecting 520 representative isolates I have trimmed large single centre studies to a single isolate; however, some geographies remained overrepresented. The other 126 isolates in **Chapter 5** came from single Thai study (188) and were chosen specifically to allow detection of bias due to dataset construction. Finally, in **Chapter 5** I used all Kp isolates collected from liver to improve the statistical power of analysis, but these isolates were not well diversified with 26 (33%) collected in Singapore. The ideal dataset would have consisted of a random selection of hypervirulent isolates either nationally or globally with matching non-hypervirulent Kp samples from the same time and geography.

Chapter 3: Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	2004066	**Title**	Mr
First Name(s)	Anton		
Surname/Family Name	Spadar		
Thesis Title	Understanding the genetic diversity, antimicrobial resistance, and virulence of Klebsiella pneumoniae bacteria		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Scientific Reports		
When was the work published?	Mar 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I have performed bioinformatic and statistical analyses, interpreted results, wrote the first draft and compiled the final manuscript

SECTION E

Student Signature	
Date	26/09/22

Supervisor Signature	
Date	26/09/22

scientific reports

OPEN

Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals

Anton Spadar[1], João Perdigão[2], Jody Phelan[1], James Charleston[1], Ana Modesto[2], Rita Elias[2], Paola Florez de Sessions[3], Martin L. Hibberd[1], Susana Campino[1], Aida Duarte[4,5] & Taane G. Clark[1,6,7]✉

Klebsiella pneumoniae **is an important nosocomial infectious agent with a high antimicrobial resistance (AMR) burden. The application of long read sequencing technologies is providing insights into bacterial chromosomal and putative extra-chromosomal genetic elements (PEGEs) associated with AMR, but also epigenetic DNA methylation, which is thought to play a role in cleavage of foreign DNA and expression regulation. Here, we apply the PacBio sequencing platform to eight Portuguese hospital isolates, including one carbapenemase producing isolate, to identify methylation motifs. The resulting assembled chromosomes were between 5.2 and 5.5Mbp in length, and twenty-six PEGEs were found. Four of our eight samples carry $bla_{CTX\text{-}M\text{-}15}$, a dominant Extended Spectrum Beta Lactamase in Europe. We identified methylation motifs that control Restriction–Modification systems, including GATC of the DNA adenine methylase (Dam), which methylates N6-methyladenine (m6A) across all our *K. pneumoniae* assemblies. There was a consistent lack of methylation by Dam of the GATC motif downstream of two genes: *fosA*, a locus associated with low level fosfomycin resistance, and *tnpB* transposase on IncFIB(K) plasmids. Overall, we have constructed eight high quality reference genomes of *K. pneumoniae*, with insights into horizontal gene transfer and methylation m6A motifs.**

Klebsiella pneumoniae (*Kp*) are Gram-negative bacteria that are found in the normal flora of the mouth, intestines, skin and faeces, but in other parts of the body, such as the lungs, can cause severe morbidity with a diverse disease spectrum that can culminate in complicated invasive infections. This pathogen is increasingly recognized as an important etiological agent of healthcare associated infections. *Kp* has also been identified as a key route of introduction and dissemination of antimicrobial resistance (AMR) genes into other clinically significant pathogens[1,2]. In Europe, *Kp* with resistance to fluoroquinolones and carbapenems, and third-generation cephalosporins has been increasing, leading to reduced treatment options[3].

The genome sequencing of *Kp* clinical isolates can provide insights into AMR, but also epigenetic information superimposed over nucleotide sequences[4]. The formation of epigenetic lineages enables the adaptation of bacterial populations to harsh or changing environments and modulates the interaction of pathogens with their eukaryotic hosts[5–7]. Epigenetic signals control DNA–protein interactions and can cause phenotypic change in the absence of mutation[8]. A common mechanism of epigenetic signalling is DNA methylation by orphan methyltransferases (MTases) such as Dam, which have roles in chromosome replication and segregation, nucleoid organization, cell cycle control, and DNA repair[4,8,9]. DNA methylation is also the key element of Restriction–Modification (R–M) systems that not only provide defence against foreign DNA, but also encourage bacterial evolution by driving the persistence of plasmids and other mobile genetic elements[10,11].

Single Molecule, Real-Time (SMRT) platforms detect DNA modifications by measuring variation in the polymerase kinetics of DNA base incorporation during sequencing. The approach has the ability to detect genome-wide MTase N6-methyladenine (m6A) and N4-methylcytosine (m4C) target motifs at coverage levels recommended for assembly, and reveal phase variation of related genes[12]. DNA adenine methylase (Dam),

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, UK. [2]Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Lisboa, Portugal. [3]Genome Institute Singapore, Singapore, Singapore. [4]Faculdade de Farmácia, Universidade de Lisboa, Lisboa, Portugal. [5]Centro de Investigação Interdisciplinar Egas Moniz, Instituto Universitário Egas Moniz, Almada, Portugal. [6]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [7]Department of Infection Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. ✉email: taane.clark@lshtm.ac.uk

ID	Hospital	Isolation Year	Fold coverage	N50 Mbps	Busco** complete / fragmented (%)	Non-chromosomal Contigs > 1kbps	MLST-type	K locus type	O locus type
Kp1363	HSAC	2005	197	5.34	89.4/7.3	1	76***	KL10	O3/O3a
Kp1208	HSM*	2006	130	5.37	98.5/0.2	2	35	KL22	O1v1
Kp1264	HSM	2007	131	5.23	97.5/0.9	4	147	KL64	O2v1
Kp1675	HSAC	2008	428	4.02	98.9/0.2	4	48	KL62	O1v1
Kp2209	HSM*	2008	504	5.37	98.5/0.5	4	133	KL116	O1v1
Kp2564	HSM	2009	36	5.39	98.7/0.2	6	11	KL111	O3b
Kp2958	HSM	2010	187	5.48	97.5/1.1	3	14	KL2	O1v1
Kp3860	HSM	2013	241	5.39	97.3/1.1	2	307	KL102	O2v1

Table 1. *Klebsiella pneumoniae* assemblies generated for the analysis. All isolates sourced from blood, except * from urine; HSM Hospital de Santa Maria; HSAC Hospital St Antonio dos Capuchos; MLST sequence type; ** Busco score is based on genes set enterobacterales_odb10; *** Kp1363 matches six out of seven ST76 alleles; N50 is defined as the sequence length of the shortest contig at 50% of the total genome length.

which methylates m6A in the GATC sequence, plays a key role in DNA mismatch repair, as well as in bacterial virulence and gene expression, including in some strains of *Kp*[10,13–15]. R–M systems have also been observed on bacterial plasmids where they may contribute to their maintenance[11]. *Kp* is known to have type I and type II R–M systems[16]. The two systems have different mechanics and their distinct motif types have been reviewed elsewhere[4,8]. Briefly, type I R–M systems consist of specificity, modification and restriction subunits. The first two subunits are usually located together on a chromosome and are under the control of the same promoter[17]. The restriction subunit in the type I R–M system recognises long bipartite motifs, such as $GGCAN_8TCG$. While not part of complete type II R–M, Dam is the most common MTase in Gamma-proteobacteria and it recognises palindromic 5′-GATC-3′ motifs[4,8].

Here we applied SMRT sequencing to eight *Kp* isolates from Portugal with antibiotic susceptibility phenotyping to characterise the bacterial epigenome and explore the relationship between methylation and AMR. We focused on methylation, including around AMR genes, and on differences in the abundance of R–M recognition motifs on *Kp* chromosomal and mobile genetic elements. The abundance of some target methylation motifs was different between chromosomes and plasmids, especially the GATC motifs methylated by orphan MTase Dam. We also found that a GATC motif immediately downstream of the *fosA* gene, which confers low level fosfomycin resistance[18], is consistently unmethylated in our samples. Isolates that had the *tnpB* transposase gene[19] on the IncFIB(K) plasmid also consistently lacked methylation immediately downstream of this gene.

Results

Genome assemblies and phylogeny.

Eight *Kp* isolates with different multi-locus sequence types (MLSTs)[20] were sourced from two hospitals in Lisbon, Portugal, between 2005 and 2013, and sequenced on the PacBio RSII technology (Table 1). The assembled chromosomes were between 5.2 and 5.5 Mbp in length and had GC content values between 57.2 and 57.7%. By contrast, 26 putative extra-chromosomal genetic elements (PEGEs) ranged in length between 3.6 and 284.2 Kbp and did not segregate into clusters based on sequence length. The PEGEs have a GC content between 41.4 and 54.1%, except for an outlying 9.9 Kbp plasmid with 60.1%. This outlying plasmid had 100% coverage and 93% identity to several plasmids of Gram-negative bacteria (e.g. CP027616.1, CP023430.1) (Table 1). To put our samples in a broader context we constructed a maximum likelihood phylogenetic tree based on the alignment of seven MLST informative and nineteen core genome loci, which contained 83 representative *Kp* MLST groups sourced from the NCBI database (Fig. 1)[21]. All our isolates were the nearest neighbour of isolates with the same MLST, but there was varying heterogeneity in genetic distance between the samples from same MLST, resulting from using 19 additional loci that could differentiate geographical and temporal differences. For example, the nearest neighbour to our Kp3860 isolate (ST307) was one collected in Malta in the same year with an identical MLST and no sequence differences across the genes analysed. On the other hand, our Kp2209 isolate had some divergence from another ST133 sample isolated in Thailand eight years earlier.

Antimicrobial resistance loci.

Consistent with their Portuguese origin, four of our eight samples carry $bla_{CTX\text{-}M\text{-}15}$, a dominant extra-spectrum beta-lactamase in Europe and a source of resistance to third-generation cephalosporins[22]. Two isolates, Kp1675 and Kp1264, carry the gene on a chromosomal region flanked by Tn2 (Tn3 family) and IS26 (family IS6) mobile genetic elements. Two further isolates, Kp3860 and Kp2209, carry $bla_{CTX\text{-}M\text{-}15}$ gene on IncFIB(K) plasmids. Finally, in addition to chromosomal $bla_{CTX\text{-}M\text{-}15}$, Kp1264 also carries it on IncFIA(HI1) plasmid which belongs to the same incompatibility group as IncFIB(K)[23]. All five $bla_{CTX\text{-}M\text{-}15}$ fragments (4 isolates) are embedded in the 9500 nt sequence that is near identical between samples (> 99.98% similarity), though Kp2209 has only 50% coverage compared to 93% coverage for the other sequences. This 9500 nt sequence also carries quinolone and aminoglycoside resistance genes. In three isolates (Kp1675, Kp1264 and Kp3860), the region containing $bla_{CTX\text{-}M\text{-}15}$ also contains $bla_{OXA\text{-}1}$, which is a penicillinase and a major correlate of resistance to piperacillin/tazobactam and co-amoxiclav in *Kp* and *E. coli*, and is commonly associated with co-carriage of *aac(6')-Ib-cr*, which restricts aminoglycoside and fluoroquinolone treatment options[24,25].

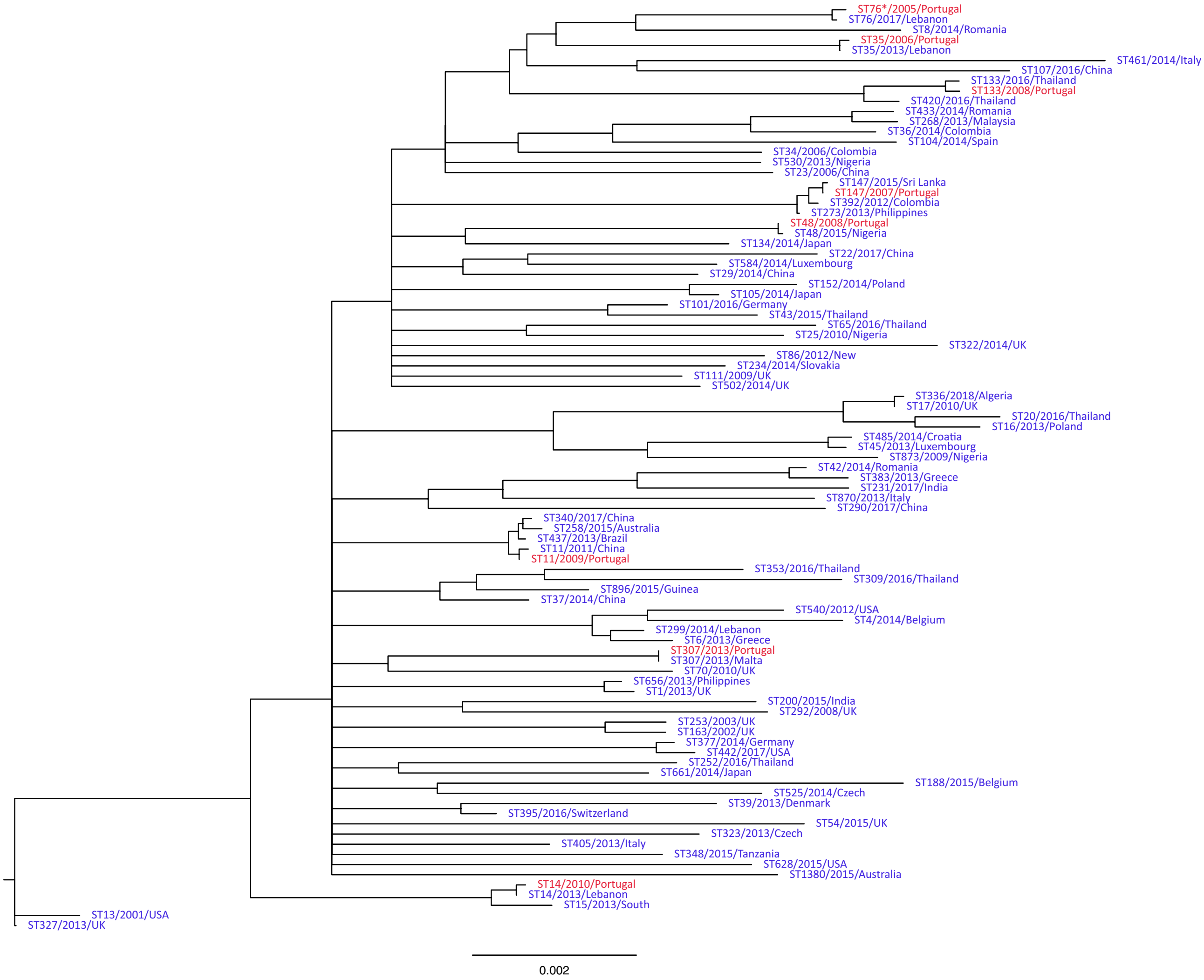


Figure 1. Unrooted maximum likelihood phylogenetic tree of common 83 *Kp* sequence types (blue) and the 8 isolates from our study (red). Edges with bootstrap support below 75 (based on 5000 bootstraps) were collapsed. Scale is nucleotide substitutions per site. Tip labels are sequence type, collection year and country.

Only the Kp2564 isolate bore a KPC-3 carbapenemase coding gene (*bla*$_{\text{KPC-3}}$), which is located on an IncFIA plasmid. This carbapenemase is thought to have originated in clonal group 258 (CG258), to which this sample belongs[26]. KPC-3 has now spread across *Kp* MLSTs globally, including in Portugal[27], where its prevalence has increased dramatically over the last decade[28]. All samples were tested for susceptibility to major anti-microbials (Table S1). Sample Kp2564 was the only one resistant to imipenem, which is consistent with presence of *bla*$_{\text{KPC-3}}$. Of all samples only Kp1264 appeared resistant to fosfomycin.

Methylation analysis. Seven of the eight samples had sufficient quality and coverage to perform DNA methylation analysis using the SMRT portal. For the other sample (Kp2564), we successfully identified methylation motifs, but low sequencing coverage meant that for non-GATC motifs the detection of motifs methylation state was unreliable. After comparing the generality of motifs versus the frequency of motif methylation (Figure S1) for each isolate, we removed the motifs for which less than 60% of motif occurrences were methylated. For Kp2564 we focused only on GATC, AACN$_5$RTGC and GCAYN$_5$GTT motifs, the latter two being part of same type I bipartite recognition sequence. While our raw results did contain four m4C modifications, the highest modification rate was 17% and therefore omitted from further analysis.

All seven high quality assemblies contained a GATC recognition motif of the Dam based type R–M system. A total of ten type I R–M system motifs were found. For each of these ten we identified specific MTases (Table 2) by comparing their specificity subunits to those in the REBASE database[16]. For all but one specificity subunit, REBASE has 100% matches of the DNA target recognition domain. The recognition sequence motif on plasmid_5 of Kp2564 was the exception with 82% highest amino acid identity in the target domain. For Kp1363, SMRT Analysis returned four motifs (CGACCN$_4$TGG, CGAYDN$_4$TGG, CCAN$_4$GATCA and CCAN$_5$RTCG) without partner motif strings. The first and second motifs are imperfect reverse complements of the fourth. The third motif is likely to be recognised by the same specificity subunit, but may show low methylation (60%) due to the

Assembly	Main motif/partner motif	Modified fraction	Number of motifs	Specificity subunit
Kp1208	GATC	98%	63,002	
Kp1208	TGAYN6TTTG/ CAAAN6RTCA	96%/88%	463	chr: 557,963–561,058
Kp1264	CCAGN7RTTC/ GAAYN7CTGG	98%/94%	347	chr: 1,565,436–1,567,199
Kp1264	GATC	98%	61,076	
Kp1264	GGCAN8TCG/ CGAN8TGCC	98%/91%	1042	chr: 1,638,656–1,639,996
Kp1363	CCAN4GATCA	60%	447	plasmid_1: 100,843–102,264
Kp1363	CCAN5RTCG	98%	2313	plasmid_1: 100,843–102,264
Kp1363	CGACCN4TGG	98%	276	plasmid_1: 100,843–102,264
Kp1363	CGAYDN4TGG	98%	1613	plasmid_1: 100,843–102,264
Kp1363	GAAAYN8TCG/ CGAN8RTTTC	97%/96%	459	chr: 1,946,700–1,947,668
Kp1363	GATC	99%	61,266	
Kp1675	GATC	99%	64,248	
Kp2209	GATC	99%	64,264	
Kp2209	GATGN6TTG/ CAAN6CATC	99%/99%	1029	plasmid_1: 2570–3844
Kp2564	AACN5RTGC/ GCAYN5GTT	26%/26%	998	plasmid_3: 94,898–96,112
Kp2564	GATC	36%	63,282	
Kp2958	ACAN8TGAC/ GTCAN8TGT	98%/96%	319	chr: 4,378,992–4,379,303
Kp2958	AGCN5CTTC/GAAGN5GCT	100%/98%	1024	chr: 1,411,097–1,412,662
Kp2958	GATC	98%	64,066	
Kp3860	GAAAN6GGG	97%	586	chr: 1,856,401–1,857,603
Kp3860	GATC	99%	63,388	

Table 2. List of high-quality *m6A* motifs. Only the GATC motif was present in multiple isolates, consistent with widespread presence of DNA adenine methylase among Gammaproteobacteria. Kp2564 has low methylation rates due to low sequencing coverage.

presence of a type II GATC motif within the type I motif. Given the highly similar number of sites (2336 for the first three motifs versus 2311 for the fourth), the similarity of motifs and only two specificity proteins in the assembly, we suspect that the four motifs are recognised by the same specificity subunit. A problematic sequence was GAAAN$_6$GGG in Kp3860 for which the SMRT Analysis pipeline did not return a partner motif, nor plausible reverse complement. The assembly has one specificity protein, but has no exact match to known recognition sequences in REBASE, with the closest being GAGN$_6$GGG from an *E. coli* specificity subunit (S.Eco77I), with identity 36% (e-value 7e-72).

Due to a lack of a corresponding restriction enzyme in *Kp*, the GATC motif is not involved in the defence against foreign DNA[8]. To assess if this is reflected in the number of motifs on the *Kp* chromosome and PEGEs, we examined the relative abundance of GATC motifs, as measured by the ratio of total length of observed motifs to those of genomic sequence. For the GATC motif, there was much greater abundance on the chromosomes compared to PEGEs (Wilcoxon $P < 3 \times 10^{-6}$) (Table S2, Fig. 2a). We found the same result in 673 high quality *Kp* assemblies from the NCBI RefSeq database (all N50s > 4.5Mbp; Fig. 2b), where the GATC abundance on the chromosome (mean 2.25%; standard deviation 0.017%) was greater than on PEGEs (mean 1.52%; std. dev. 0.088%) (Wilcoxon $P < 0.001$). We also investigated the abundance of type I R–M system motifs (i.e. non-GATC) by comparing the share of chromosomes occupied by the motifs that are recognised by assembly's MTases versus the share occupied by recognition sequences from other assemblies. There was no strong difference in the abundance of type I recognition motifs on chromosomes or between chromosomes and PEGEs (Wilcoxon $P > 0.230$). Similarly, there was no strong difference between PEGEs that were recognised by their own assembly's R–M system compared to those that were not recognised (Wilcoxon $P = 0.187$).

Unmethylated motifs. Methylated motifs (see Table 2) covered a large portion of the genome and were relatively evenly distributed. Therefore, we focused on unmethylated motifs 50bps upstream and downstream of the genes. Upstream sequences capture gene promoters, whereas downstream sequences provide a control. We focused on the genes and motifs that have higher than expected number of unmethylated motifs (Table 3) in most isolates (> 3/7) (see Data S1 for the full list of identified genes). All such cases had a GATC motif, with only *fosA* and *tnpB* IS3 family (IS2 group) transposase genes having a consistently unmethylated motif downstream. *TnpB* is a key component of mobile genetic elements' transposition mechanism, and in four samples that had the gene, it was located on a large (197–284kbp) IncFIB(B) plasmid containing various heavy metal and AMR loci. The *fosA* gene provides *Kp* with an inherent low-level resistance to the fosfomycin antibiotic[29,30]. Only one isolate (Kp1264) was resistant to fosfomycin (Table S1), but it was one of two isolates (Kp1264 and Kp1363; Fig. 3) that had the R–M system inserted after *fosA*. Both *fosA* and *tnpB* genes lacked methylation at a GATC motif downstream of the gene, so the unmethylated GATC should not be directly affecting promoter region; however, the lack of methylation may indicate that MTase cannot access the site.

Of the seven isolates with high quality site methylation data, only Kp1264 (ST147) and Kp1363 (ST76) did not have the unmethylated motif downstream of *fosA*. In both isolates, the required GATC motif was missing

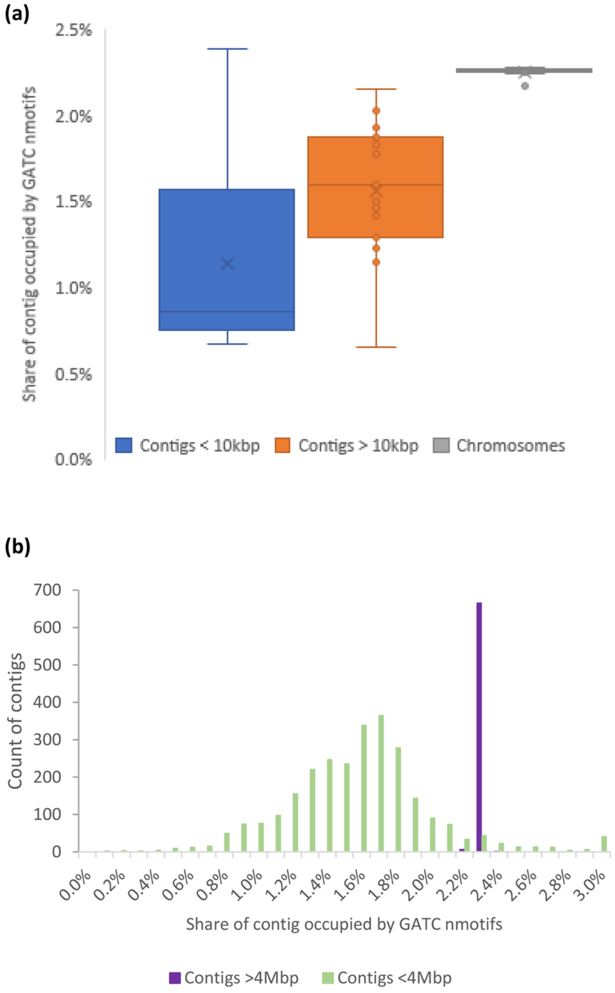


Figure 2. (**a**) The share of genomic elements occupied by a GATC motif in the assemblies in Table 1. The largest non-chromosome contig, 284kbp, has abundance ratio of 1.8%. (**b**) GATC abundance in 673 high quality assemblies of *Kp* taken from NCBI database limiting sequences to those that have N50 in excess of 5Mbp. The mean chromosomal GATC abundance in (**a**) and (**b**) is the same.

Location	Product	RefSeq	GATC motif 50 bps up/down of gene is unmethylated							Binomial test p-value
			Kp1675	Kp1264	Kp1208	Kp1363	Kp2958	Kp2209	Kp3860	
Up	D-ribose transporter subunit RbsB	VEC00318.1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	1.8E-12
Up	Branched-chain amino acid transport protein azlC	AHM82117.1	Yes	Yes	Yes	NM	Yes	Yes	Yes	8.6E-11
Up	D-arabinitol dehydrogenase	BAH64332.1	No	Yes	Yes	Yes	Yes	Yes	Yes	5.9E-10
Up	DnaK suppressor protein	BAH61791.1	Yes	No	Yes	Yes	Yes	Yes	Yes	5.9E-10
Up	BCCT family transporter	AHM80284.1	NM[a]	Yes	Yes	Yes	Yes	Yes	No	2.4E-08
Down	FosA family fosfomycin resistance glutathione transferase	QBH08895.1	Yes	NM	Yes	NM	Yes	Yes	Yes	4.1E-09
Down	IS3 family transposase	VEC38624.1	NM	NM	Yes	NM	Yes	Yes	Yes	1.9E-07
Up	Transcriptional regulators of sugar metabolism	VEC00015.1	No	Yes	Yes	NM	Yes	No	Yes	2.8E-06
Down	DUF1145 family protein	AHM77076.1	Yes	No	Yes	Yes	No	Yes	No	6.4E-06

Table 3. Genes which have an unmethylated GATC motif 50bps upstream/downstream in at least four samples. Every sample has either one copy of the gene or no copies. The p-value null hypothesis is probability of gene's GATC motif being unmethylated is the same as probability of any GATC motif being unmethylated. None of the non-GATC motifs had a p-value below 0.05. NM = no motif. [a]Missing start codon.

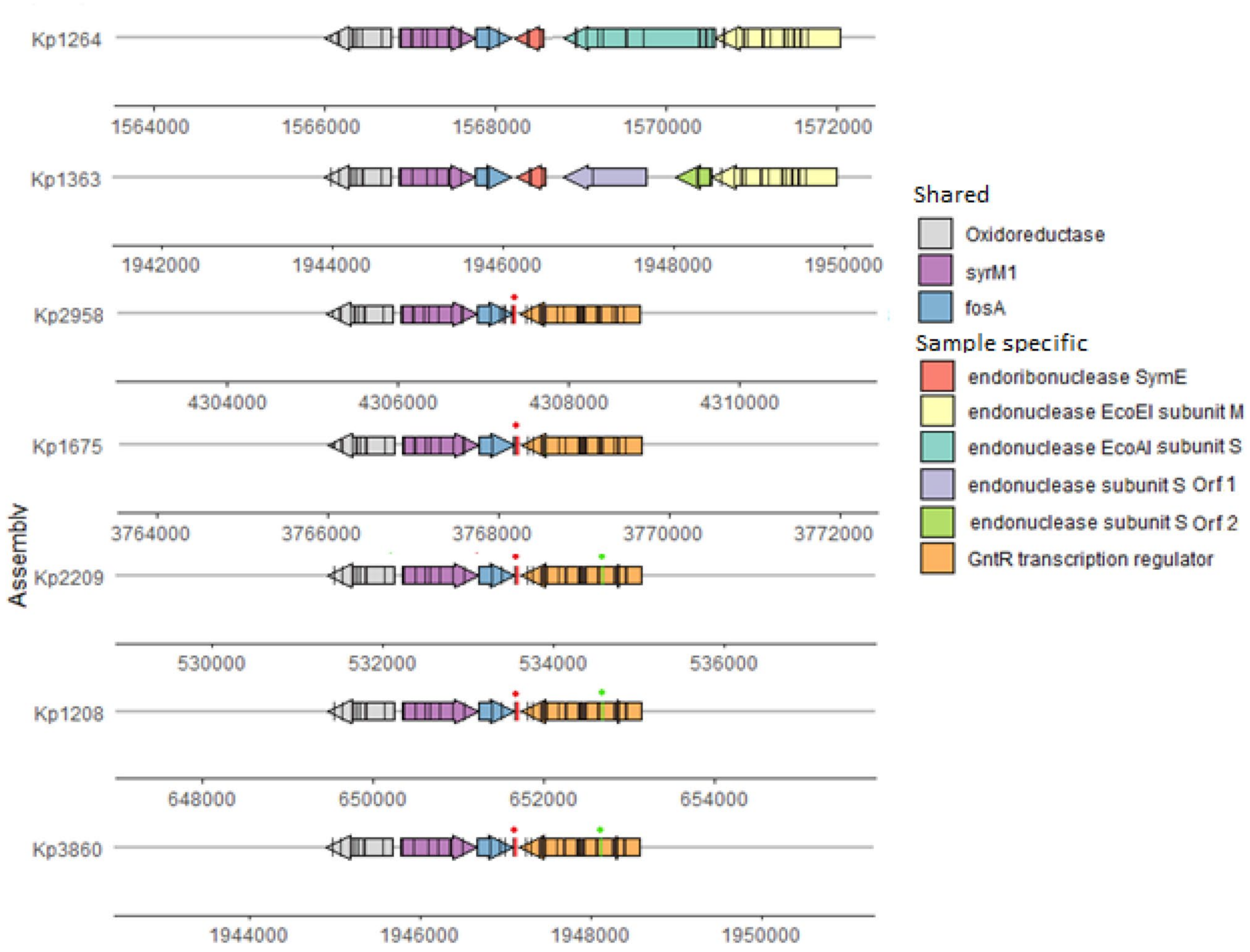


Figure 3. The *fosA* gene and surrounding region. (Black bars) methylated GATC motifs; (red bars with red dot above) unmethylated GATC motifs downstream of *fosA* on the same strand; (green bars with green dot above) unmethylated GATC on the opposite strand of GntR family transcription regulator coding sequence. Kp1264 and Kp1363 lack a GATC motif downstream of *fosA* due to an insertion of endonuclease. The other samples all have a single unmethylated GATC motif downstream of *fosA* on the same strand as *fosA* coding sequence. The only other unmethylated GATC in these regions are the motifs on GntR family transcription regulator.

Gene or product	ST258	ST11	ST15	ST147	ST405	ST101	ST231	Other	% of Total
GntR family transcriptional regulator	209	477	200				57	1408	68.3
Putative protein			1					306	8.6
DNA-dependent helicase II					23			289	8.1
5′-nucleotidase						88		130	6.1
Toxic protein SymE				130				56	5.2
endoribonuclease SymE								59	1.7
yjiR_1	1	3						18	0.7
yjiR_2	8	3	1					5	0.5
Hypothetical protein	1					6		7	0.4
Type I restriction endonuclease								6	0.2

Table 4. Genes and proteins encoded by coding sequences immediately downstream on opposite strand of chromosomal *fosA*. Together these account for 99.6% of downstream elements in 3584 *K. pneumoniae* samples. For a given MLST, all samples normally have the same gene downstream of *fosA*. Genes *yjiR_1* and *yjiR_2* are truncated versions of the GntR family transcription regulator.

due to insertion of an endonuclease gene (> 3 kbp) (Fig. 3). To assess the frequency of this insertion we analysed available *Kp* assemblies (n = 3584) from the NCBI database. We found that 99.4% of assemblies had the *syrM1* gene upstream of chromosomal *fosA*. The exceptions were 23 samples from diverse geographical regions and sequence types (Table S3). Further, we observed a limited variety of genes downstream on the opposite strand of *fosA* (Table 4). The top five genes accounted for 96.2% of the assemblies, with the predominant one being the GntR family transcription regulator (68.3%)[31]. For the ST147 sequence type (NCBI, n = 131), we found most assemblies (130/131) had toxic protein SymE in the same position. The outlying assembly (GCA_004145685.1_ASM414568v1_genomic) had a truncated version of SymE due to the scaffold ending in the coding sequence. For the ST76 sequence type, all NCBI isolates (n = 7) and Kp1363 (one locus mismatch) had coding sequence of endoribonuclease SymE downstream of *fosA* (Table 4).

For a random subset of unmethylated motifs (from Table 3), we examined the IPDs of individual reads to evaluate the existence of distinct subpopulations with regard to the methylation status of genes, thereby investigating potential intra-populational epigenetic diversification (Figure S4). Standard SMRT Analysis output files report the methylated fraction of reads in the sample, but only for motifs that were called as methylated. We modified the SMRT Analysis algorithms to produce per base IPD values for all nucleotides. After analysing the underlying reads, we found that IPD levels in some genes, including *fosA* and *dksA* (DnaK suppressor protein), differentiated two subpopulations corresponding to methylated and unmethylated cells. By contrast, we observed no methylated subpopulation in *mglB* (D-ribose transporter subunit RbsB).

Discussion

Across seven *Kp* isolates with high quality PacBio data, we identified the common Dam based methylation mechanism involving GATC motifs, as well as several type I R–M systems. With a type I R–M system, we observed that the abundance of recognition motifs did not vary between chromosomes of different assemblies, but that some depletion of recognition motifs has occurred in PEGEs. This observation supports the role of type I R–M system as a defence mechanism against invasion by foreign DNA, but not its regulatory function, as chromosomal abundance was not correlated to specificity subunit. In contrast, there was a clear difference in abundance of the GATC motif on chromosomes and PEGEs, which suggests that the motif has a function as identified by previous research[8,9].

Because GATC motifs are known to have a regulatory function, we expected to find that some genes had consistently unmethylated GATC in gene promoter regions[8,9]. However, the finding of consistently unmethylated GATC downstream of genes (e.g. *fosA* and *tnpB*) is intriguing. This outcome may be an indication of a secondary structure that is inaccessible to MTase, or that the GATC motif downstream of *fosA* and *tnpB* is a distant promoter or regulatory region. The phenotypic impact of this GATC motif requires follow-up experiments ideally via single nucleotide site-directed mutagenesis as to prevent methylation with minimum downstream effects. Analysis of this region has also led us to identify the limited variety of recombination events around *fosA* and their potential MLST specificity. Toxic protein SymE is part of the plasmid toxin-antitoxin system. To assess its prevalence in *Kp* we evaluated 131 randomly chosen ST147 samples from the NCBI database. We found the same *fosA*-*SymE* sequence in all but one of them, but in no other major ST. The potential impact of the *SymE* insertion on fosfomycin resistance should be evaluated in follow-up experiments. The other coding sequences in that location do not appear to be part of the toxin-antitoxin system[32], and therefore this locus has the potential to evaluate the robustness of a *Kp* phylogeny based on sequence types. We would expect that sequence types with the same insertion would cluster together, and the opposite observation would indicate possible recombination of MLST genes.

The general lack of resistance to fosfomycin in *Kp* species, despite decades of active use of fosfomycin for infection control[30,33], is in contrast to the emergence of resistance to other antimicrobials. However, *Kp* has been reported to acquire resistance *in vitro* within 24 hours of exposure to fosfomycin[34]; though this resistance does not seem to persist as clinical studies report the limited spread of fosfomycin resistance[33]. The identified unmeth-ylated GATC motif downstream of *fosA* may be correlated with the rapid acquisition of fosfomycin resistance in individual isolates. Our dataset is too small to draw any robust statistical inference, but the only sample in our study resistant to fosfomycin, Kp1264, is one of the two samples which lacks a GATC motif downstream of *fosA* (Fig. 3). The plasmid toxin-antitoxin system downstream of *fosA* in both Kp1363 and Kp1264 requires constant transcription, which means the genomic region is accessible to transcription machinery; whereas unmethyl-ated GATC may be the result of DNA regional conformation inaccessible to Dam. It is unlikely that a lack of methylation downstream of the *fosA* gene is itself the fosfomycin resistance mechanism. If correlation exists, the more likely explanation is that conformation of the genomic region that prevents methylation, also pre-vents *fosA* transcription. This hypothesis is particularly relevant for the treatment of complicated infections by carbapenemase-producing strains, for which fosfomycin is often used in combination with another drug (e.g. colistin) as a last resort therapeutic option[35,36].

Overall, our work has provided new insights into methylation and potential MLST specificity. We have generated eight new reference genomes for *Kp* and analysed their methylomes. The findings reinforce the role of type I motifs as a defence mechanism against foreign DNA. We also identified a higher than expected rates of the GATC motif on *Kp* chromosomes which, together with absence of respective endonuclease, support existence of Dam function such as DNA mismatch repair[37] and gene expression regulation[9]. We identified two genes, *fosA* and *tnpB,* which have an unmethylated GATC motif downstream of each locus. Methylation analysis may be useful for identification of distant regulatory regions or frequent secondary DNA structures. In particular, a lack of methylation downstream of *fosA* genes could explain not only rapid emergence of resistance in individual samples, but also lack of widespread resistance in *Kp*. Further, using a bioinformatics approach we detected the presence of both methylated and unmethylated sequencing reads existing within individual samples, potentially representing subpopulations of distinct epigenetic lineages, which could contribute to stochastic phenotypic switching mechanisms in bacteria[38,39]. Such heterogeneity can be investigated in more depth using a SMALR approach[40]. This important biology and other findings should be evaluated in larger genomic and functional studies, and could lead to new insights into *Kp* infection control.

Methods

Sample collection, culture and sequencing. All eight samples were collected from Hospital de Santa Maria (n = 6) and Hospital St Antonio dos Capuchos (n = 2) in Lisbon, Portugal. The samples were cultured and tested for AMR as described previously[28]. DNA was extracted from strain cultures, grown overnight at 37°C on Mueller–Hinton Agar. DNA extraction was carried out using the Cetyl trimethylammonium bromide (CTAB) method previously described using Tris–Acetate buffer (10 mM, pH 8)[41]. Our samples were generated without TET1 oxidation; thus we could not examine 5mC methylation[12,42]. All DNA samples were sequenced at the Genome Institute Singapore on the PacBio RSII platform.

PCR and Sanger sequencing-based MLST analysis was based on fragments of seven housekeeping genes: *rpoB* (beta-subunit of RNA polymerase), *gapA* (glyceraldehyde 3-phosphate dehydrogenase), *mdh* (malate dehydrogenase), *pgi* (phosphoglucoseisomerase), *phoE* (phosphorine E), *infB* (translation initiation factor 2), and *tonB* (periplasmic energy transducer). Details of the MLST scheme including amplification and sequencing primers, allele sequences and MLSTs are available on the Institute Pasteur's MLST Web site (https://bigsdb.pasteur.fr/klebsiella/klebsiella.html).

Bioinformatic analysis. A summary of the bioinformatic and analysis pipeline is provided (Figure S3). Data assembly was performed using three different software tools (HGAP3, Canu and Flye). HGAP3 is part of the SMRT Analysis toolkit (version 2.3), while Canu and Flye software are standalone[43–45]. Short read polishing was not performed due to > 100-fold long read coverage for each sample. We have aligned raw reads to the assemblies and examined region ± 2000 nt around *fosA* gene in each assembly. We did not observe[46] any SNPs, InDels or alignment abnormalities. Between samples, the coverage of the region ranged from 77- to 219-fold. Generated assemblies were compared using the Busco enterobacteriales (odb10) dataset[47]. For each sample the assembly with highest Busco complete gene count was selected. The sum of complete and fragmented genes yielded the same "best" assemblies. While the HGAP3 assembly had marginally higher N50 score, in most cases its Busco scores were substantially lower. All but one of the selected assemblies had a scaffold longer than 5Mbp. The outlying assembly, Kp1675, had a chromosome split into two scaffolds which was considered an acceptable trade-off for a 5.3% higher share of complete genes. The *in silico* resistance profile, MLST, O-locus and K-locus types were determined using the kleborate tool[48,49]. Plasmid and Insertion Sequence (IS) identities were estab-lished using PlasmidFinder[50] and ISFinder[51].

Methylation analysis. We used the generated assemblies to perform methylation detection by applying the PacBio SMRT Analysis toolkit (v2.3). To discriminate between natural variation in IPD and true methylation the toolkit relies on expected distribution of unmethylated IPDs given the nucleotide sequence context. When IPD value exceeds the probability threshold, we used default p-value 0.01, the relevant base is called as modified. The calculations are implemented in *KineticWorker.py*, which is part of SMRT Analysis toolkit[52]. After comparing the generality of motifs versus the frequency of motif methylation for each isolate, we removed the motifs for which less than 60% of motif occurrences were methylated. We retained two motifs for the Kp2564 isolate which had methylation fractions below 60% (Table 2), where the low level of methylation was result of insufficient sequencing coverage.

NCBI datasets. To assess the abundance of GATC motif on *Kp* chromosomes, we downloaded from the NCBI RefSeq database all *Kp* assemblies where their longest chromosomal contig was at least 4.5Mbps (n = 673)[53]. For an analysis of the prevalence of GATC motif downstream of *fosA* gene, we downloaded all *Kp* assemblies in the NCBI pathogen database (as of November 15, 2019; n = 3584)[21]. For reconstruction of a *Kp* phylogenetic tree we determined the MLST[20] of each of the 3584 assemblies using kleborate software[48]. For each MLST which had more than nine samples from different geographical locations (city, region, or country) and year of collection data, we included a single randomly selected assembly for phylogenetic reconstruction. A total of 83 isolates were used in the construction of the phylogenetic tree. The tree was constructed using IQTREE software[54] with the GTR + F + G4 model.

Ethical statement. The study was approved by the Hospital de Santa Maria and Hospital St Antonio dos Capuchos Ethics Committees, and all methods were performed in accordance with the relevant guidelines and regulations, including informed consent from all patients.

Data availability

The assembled sequences are deposited in the European Nucleotide Archive (project PRJEB38289). Accession numbers and metadata are presented in Table S1. The authors confirm all supporting data, code and protocols have been provided within the article or through supplementary files.

Received: 11 September 2020; Accepted: 11 February 2021
Published online: 22 March 2021

References

1. 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).
2. Navon-Venezia, S., Kondratyeva, K. & Carattoli, A. Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol. Rev.* **41**, 252–275 (2017).
3. European Centre for Disease Prevention and Control. *Surveillance of antimicrobial resistance in Europe 2018*. (2018).
4. Blow, M. J. *et al.* The epigenomic landscape of prokaryotes. *PLOS Genet.* **12**, e1005854 (2016).
5. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61–69 (2018).
6. Phelan, J. *et al.* Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. *Sci. Rep.* **8**(1), 160. https://doi.org/10.1038/s41598-017-18188-y (2018).
7. Estibariz, I. *et al.* The core genome m5C methyltransferase JHP1050 (M.Hpy99III) plays an important role in orchestrating gene expression in *Helicobacter pylori*. *Nucleic Acids Res.* **47**, 2336–2348 (2019).
8. Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: From the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).
9. Adhikari, S. & Curtis, P. D. DNA methyltransferases and epigenetic regulation in bacteria. *FEMS Microbiol. Rev.* **40**, 575–591 (2016).
10. Nye, T. M. *et al.* DNA methylation from a type I restriction modification system influences gene expression and virulence in streptococcus pyogenes. *PLoS Pathog.* **15**(6), e1007841. https://doi.org/10.1371/journal.ppat.1007841 (2019).
11. Wang, R., Lou, J. & Li, J. A mobile restriction modification system consisting of methylases on the IncA/C plasmid. *Mob. DNA* **10**, 26 (2019).
12. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* **20**, 157–172 (2019).
13. Casselli, T. *et al.* DNA methylation by restriction modification systems affects the global transcriptome profile in Borrelia burg-dorferi. *J. Bacteriol.* **200**(24), e00395–18. https://doi.org/10.1128/JB.00395-18 (2018).
14. Pirone-Davies, C. *et al.* Genome-wide methylation patterns in *Salmonella enterica* subsp. enterica Serovars. *PLoS ONE* **10**(4), e0123639. https://doi.org/10.1371/journal.pone.0123639 (2015).
15. Kumar, S. *et al.* N4-cytosine DNA methylation regulates transcription and pathogenesis in *Helicobacter pylori*. *Nucleic Acids Res.* **46**, 3429–3445 (2018).
16. Roberts, R. J., Vincze, T., Posfai, J. P. & Macelis, D. REBASE: Restriction enzymes and methyltransferases. *Nucleic Acids Res.* **31**, 418–420 (2003).
17. Murray, N. E. Type I Restriction Systems: Sophisticated Molecular Machines (a Legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.* **64**, 412–434 (2000).
18. Elliott, Z. S. *et al.* The Role of fosA in challenges with fosfomycin susceptibility testing of multispecies Klebsiella pneumoniae carbapenemase-producing clinical isolates. *J. Clin. Microbiol.* **57**(10), e00634–19. https://doi.org/10.1128/JCM.00634-19 (2019).
19. Bao, W. & Jurka, J. Homologues of bacterial TnpB-IS605 are widespread in diverse eukaryotic transposable elements. *Mob. DNA* **4**, 12 (2013).
20. Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. D. & Brisse, S. Multilocus sequence typing of Klebsiella pneumoniae nosocomial isolates. *J. Clin. Microbiol.* **43**, 4178–4182 (2005).
21. Agarwala, R. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkx1095 (2018).
22. Kazmierczak, K. M., de Jonge, B. L. M., Stone, G. G. & Sahm, D. F. Longitudinal analysis of ESBL and carbapenemase carriage among Enterobacterales and Pseudomonas aeruginosa isolates collected in Europe as part of the international network for optimal resistance monitoring (INFORM) global surveillance programme, 2013–17. *J. Antimicrob. Chemother.* https://doi.org/10.1093/jac/dkz571 (2020).
23. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* **31**, e00088-e117 (2018).
24. Sugumar, M., Kumar, K. M., Manoharan, A., Anbarasu, A. & Ramaiah, S. Detection of OXA-1 β-lactamase gene of Klebsiella pneumoniae from blood stream infections (BSI) by conventional PCR and *in-silico* analysis to understand the mechanism of OXA mediated resistance. *PLoS ONE* **9**(3), e91800. https:// doi. org/ 10. 1371/ journ al. pone. 00918 00 (2014).
25. Xiang Yang Zhou, Bordon, F., Sirot, D., Kitzis, M. D. & Gutmann, L. Emergence of clinical isolates of Escherichia coli producing TEM-1 derivatives or an OXA-1 β-lactamase conferring resistance to β-lactamase inhibitors. *Antimicrob. Agents Chemother.* **38**, 1085–1089 (1994).
26. Chen, L. *et al.* Carbapenemase-producing Klebsiella pneumoniae: Molecular and genetic decoding. *Trends Microbiol.* **22**, 686–696 (2014).
27. Rodrigues, C. *et al.* KPC-3-producing Klebsiella pneumoniae in Portugal linked to previously circulating non-CG258 lineages and uncommon genetic platforms (Tn4401d-IncFIA and Tn4401d-IncN). *Front. Microbiol.* **7**, 1000. https://doi.org/10.3389/fmicb.2016.01000 (2016).
28. Perdigão, J. *et al.* Whole-genome sequencing resolves a polyclonal outbreak by extended-spectrum beta-lactam and carbapenem-resistant *Klebsiella pneumoniae* in a Portuguese tertiary-care hospital. *Microb. Genomics* https://doi.org/10.1099/mgen.0.000349(2020).
29. Ito, R. *et al.* Widespread fosfomycin resistance in gram-negative bacteria attributable to the chromosomal fosA gene. *MBio* **8**, e00749-e817 (2017).
30. Aghamali, M. *et al.* Fosfomycin: mechanisms and the increasing prevalence of resistance. *J. Med. Biol.* **68**, 11–25 (2019).
31. Suvorova, I. A., Korostelev, Y. D. & Gelfand, M. S. GntR family of bacterial transcription factors and their DNA binding motifs: Structure, positioning and co-evolution. *PLoS ONE* **10**, e0132618 (2015).
32. Shao, Y. *et al.* TADB: A web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. *Nucleic Acids Res.* **39**, D606–D611 (2011).
33. Karageorgopoulos, D. E., Wang, R., Yu, X.-H. & Falagas, M. E. Fosfomycin: evaluation of the published evidence on the emergence of antimicrobial resistance in Gram-negative pathogens. *J. Antimicrob. Chemother.* **67**, 255–268 (2012).
34. Diep, J. K., Sharma, R., Ellis-Grosse, E. J., Abboud, C. S. & Rao, G. G. Evaluation of activity and emergence of resistance of Poly-myxin B and ZTI-01 (fosfomycin for injection) against KPC-producing klebsiella pneumoniae. *Antimicrob. Agents Chemother.* **62**(2), e01815–17. https://doi.org/10.1128/AAC.01815-17 (2018).
35. Grabein, B., Graninger, W., Rodríguez Baño, J., Dinh, A. & Liesenfeld, D. B. Intravenous fosfomycin—back to the future Systematic review and meta-analysis of the clinical literature. *Clin. Microbiol. Infect.* **23**, 363–372 (2017).
36. Popovic, M., Steinort, D., Pillai, S. & Joukhadar, C. Fosfomycin: An old, new friend?. *Eur. J. Clin. Microbiol. Infect. Diseases* **29**, 127–142 (2010).
37. Marinus, M. G. DNA Mismatch Repair. *EcoSal Plus* **5**, (2012).
38. Beaumont, H. J. E., Gallie, J., Kost, C., Ferguson, G. C. & Rainey, P. B. Experimental evolution of bet hedging. *Nature* **462**, 90–93 (2009).
39. Casadesús, J. & Low, D. A. Programmed heterogeneity: Epigenetic mechanisms in bacteria. *J. Biol. Chem.* **288**, 13929–13935 (2013).

40. Beaulaurier, J. *et al.* Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nat. Commun.* **6**, 1–12 (2015).
41. Parish, T., Stoker, N. G., van Soolingen, D., de Haas, P. E. W. & Kremer, K. Restriction Fragment Length Polymorphism Typing of Mycobacteria. in *Mycobacterium Tuberculosis Protocols* 165–203 (Humana Press, 2003). http://dx.doi.org/https://doi.org/10.1385/1-59259-147-7:165
42. Clark, T. A. *et al.* Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* **11**, 4 (2013).
43. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
44. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
45. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
46. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Can. Res.* **77**, e31–e34 (2017).
47. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. in *Methods in Molecular Biology* **1962**, 227–245 (Humana Press Inc., 2019).
48. Wyres, K. L. *et al.* Identification of Klebsiella capsule synthesis loci from whole genome data. *Microb. Genomics* **2**, e000102 (2016).
49. 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**, 197–215 (2018).
50. Carattoli, A. *et al. In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).
51. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–6. https://doi.org/10.1093/nar/gkj014 (2006).
52. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
53. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
54. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

Acknowledgements

We gratefully acknowledge the Scientific Computing Group for data management and compute infrastructure at Genome Institute of Singapore for their assistance. The MRC eMedLab computing resource was used for bioinformatics and statistical analysis.

Author contributions

J.Pe., S.C., A.D., and T.G.C. conceived and directed the project. A.D. coordinated sample collection. J.P., R.E. and A.M. undertook sample processing and DNA extraction. P.Fd.S., M.L.H. and S.C. coordinated sequencing. A.S, J.Ph. and JC performed bioinformatic and statistical analyses under the supervision of S.C. and T.G.C.. A.S., J.Pe., S.C. and T.G.C. interpreted results. A.S. wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. A.S. and T.G.C. compiled the final manuscript.

Funding

S.C. is funded by Bloomsbury SET, Medical Research Council UK (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant No. BB/R013063/1) grants. T.G.C. is funded by the Medical Research Council UK (Grant Nos. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC UK (Grant No. BB/R013063/1). This study was supported in part by UID/DTP/04138/2019 from Fundação para a Ciência e Tecnologia (FCT), Portugal. J.Pe. is supported by FCT through Estímulo Individual ao Emprego Científico (CEECIND/00394/2017).

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/s41598-021-85724-2.

Correspondence and requests for materials should be addressed to T.G.C.

Reprints and permissions information is available at www.nature.com/reprints.

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

Chapter 4: Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating antimicrobial resistance

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	2004066	**Title**	Mr
First Name(s)	Anton		
Surname/Family Name	Spadar		
Thesis Title	Understanding the genetic diversity, antimicrobial resistance, and virulence of Klebsiella pneumoniae bacteria		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Scientific Reports		
When was the work published?	Aug 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	N/A		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I have analysed the data, wrote the first draft of the manuscript and edited manuscript drafts.

SECTION E

Student Signature	
Date	26/09/22

Supervisor Signature	
Date	26/09/22

scientific reports

OPEN

Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating antimicrobial resistance

Anton Spadar[1], Jody Phelan[1], Rita Elias[2], Ana Modesto[2], Cátia Caneiras[3], Cátia Marques[4], Luís Lito[5], Margarida Pinto[6], Patrícia Cavaco-Silva[7,8], Helena Ferreira[9], Constança Pomba[10], Gabriela J. Da Silva[11], Maria José Saavedra[12], José Melo-Cristino[5,13], Aida Duarte[7,14], Susana Campino[1], João Perdigão[2] & Taane G. Clark[1,15]✉

Klebsiella pneumoniae* (Kp) bacteria are an increasing threat to public health and represent one of the most concerning pathogens involved in life-threatening infections and antimicrobial resistance (AMR). To understand the epidemiology of AMR of Kp in Portugal, we analysed whole genome sequencing, susceptibility testing and other meta data on 509 isolates collected nationwide from 16 hospitals and environmental settings between years 1980 and 2019. Predominant sequence types (STs) included ST15 (n = 161, 32%), ST147 (n = 36, 7%), ST14 (n = 26, 5%) or ST13 (n = 26, 5%), while 31% of isolates belonged to STs with fewer than 10 isolates. AMR testing revealed widespread resistance to aminoglycosides, fluoroquinolones, cephalosporins and carbapenems. The most common carbapenemase gene was *bla_{KPC-3}*. Whilst the distribution of AMR linked plasmids appears uncorrelated with ST, their frequency has changed over time. Before year 2010, the dominant plasmid group was associated with the extended spectrum beta-lactamase gene *$bla_{CTX-M-15}$*, but this group appears to have been displaced by another carrying the *bla_{KPC-3}* gene. Co-carriage of *bla_{CTX-M}* and *bla_{KPC-3}* was uncommon. Our results from the largest genomics study of Kp in Portugal highlight the active transmission of strains with AMR genes and provide a baseline set of variants for future resistance monitoring and epidemiological studies.

[1]Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. [2]Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Lisbon, Portugal. [3]Microbiology Research Laboratory of Environmental Health (EnviHealthMicro Lab), Institute of Environmental Health (ISAMB) and Institute of Preventive Medicine and Public Health (IMP&SP), Faculty of Medicine, Universidade de Lisboa, Lisbon, Portugal. [4]Faculdade de Medicina Veterinária, Universidade Lusófona de Humanidades E Tecnologias, Lisbon, Portugal. [5]Laboratório de Microbiologia, Serviço de Patologia Clínica, Centro Hospitalar Universitário Lisboa Norte, Lisbon, Portugal. [6]Laboratório de Microbiologia, Serviço de Patologia Clínica, Centro Hospitalar Universitário Lisboa Central, Lisbon, Portugal. [7]Centro de Investigação Interdisciplinar Egas Moniz, Instituto Universitário Egas Moniz, Caparica, Portugal. [8]Technophage, Lisboa, Portugal. [9]UCIBIO, Microbiology Service, Biological Sciences Department, Faculty of Pharmacy, University of Porto, Porto, Portugal. [10]Centre of Interdisciplinary Research in Animal Health (CIISA), Faculty of Veterinary Medicine, University of Lisbon, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal. [11]Faculty of Pharmacy and Center for Neurosciences and Cell Biology, University of Coimbra, Coimbra, Portugal. [12]Laboratory Medical Microbiology, Department of Veterinary Sciences, CITAB-Centre for the Research and Technology Agro-Environmental and Biological Sciences, University of Trás-Os-Montes and Alto Douro, Vila Real, Portugal. [13]Instituto de Microbiologia, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal. [14]Faculdade de Farmácia, Universidade de Lisboa, Lisbon, Portugal. [15]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. ✉email: taane.clark@lshtm.ac.uk

nature portfolio

Klebsiella pneumoniae (Kp) is a common gram-negative pathogen associated with both hospital and community acquired infections[1]. During the past decades the association of Kp bacteria with antibiotic resistance has been increasingly recognized along with a variety of resistance mechanisms[2]. Production of laterally transferable genes encoding enzymes such as aminoglycoside-modifying enzymes drives high levels resistance to aminoglycoside antibiotics[3]. Resistance associated mutations in quinolone target enzymes (DNA gyrase and topoisomerase IV) are chromosomally encoded by the *gyrA, gyrB* and *parC* quinolone resistance-determining region as well as plasmid-mediated quinolone resistance genes[4,5]. Extended spectrum beta-lactamases (ESBLs) render inactive virtually all beta-lactam antibiotics (e.g., penicillins, cephalosporins, but not carbapenems). Nonetheless, of most concern is the increase in prevalence of carbapenemase producing Kp isolates[6] which greatly limits the options for effective therapies and drives hospital outbreaks that promote further spread of antimicrobial resistance (AMR)[7,8].

In Southern Europe OXA-48-like, VIM and KPC are the dominant carbapenemase families[9,10]. Knowledge of the carbapenemase landscape in Portugal is incomplete. However, the prevalence of carbapenem-resistant invasive Kp isolates in Portugal has increased from 3.4% in 2015 to 10.9% in 2019. During the same period the share of isolates with combined resistance to third generation cephalosporins, fluoroquinolones and aminoglycosides (26.5% in 2019) oscillated without an upward trend[6]. Several recent small epidemiological studies in Portugal have focused on carbapenemase producing Kp isolates[11–16], and bla_{KPC-3} was the most frequently identified carbapenem resistance gene. The bla_{KPC-3} gene is most frequently located within the Tn*4401*d transposon, and IncN, IncFII, IncFIB and IncFIIA plasmid families are the main traffickers[11,13,15,17]. High flexibility of the Kp accessory genome is an additional concern because it acts as gateway for the introduction of new resistance genes into a broader set of gram-negative pathogens[2].

Whole genome sequencing (WGS) has revolutionised the study of pathogens, not only through the characterisation of AMR associated mutations and plasmids[18,19], but also through the determination of phylogenies and transmission events[20,21]. The Kp phylogeny normally forms clades consistent with the commonly used multilocus sequence typing scheme (MLST), which is based on seven gene loci (*gapA, infB, mdh, pgi, phoE, rpoB* and *tonB*)[22]. Lipopolysaccharide (O-type) and Capsular Polysaccharide (K-type) serotype profiles can be informative for vaccine development, and > 130 capsular serotypes have been predicted from WGS data with KL2, or O1, O2, O3, and O5 serotypes accounting for most strains[23,24]. Alternatively, clonal groups (CGs) based on 694 core genes are sometimes characterised[25]. Strains from different sequence types (STs), O and K antigen types, and CGs can differ sharply in their virulence and propensity to be antibiotic resistant[25]. For example, ST23, ST26, ST57 and ST163 have been linked to Kp hypervirulence[26]. Unfortunately, the widely used Kp MLST scheme does not allow for a high-resolution phylogeny. The recombination and horizontal gene transfer within Kp also complicates a phylogenetic analysis[27], even within the same STs[28]. Furthermore, a large part of the core genome exhibits very limited polymorphism, and mobile genetic elements may have relatively greater clinical relevance as exemplified by virulence and AMR factors[27]. More generally, these include integrative conjugative elements (ICE), which are a diverse group of chromosomally integrated, self-transmissible mobile genetic elements that are active in shaping the functions of bacteria and bacterial communities, including in Kp.

The current study aims to improve understanding of the genomic and AMR landscape of Kp in Portugal by analysing the largest WGS dataset to date, which consists of 509 isolates spanning a period between years 1980 and 2019. We examine the prevalence of STs, genes associated with AMR and virulence, and how the AMR profiles relate to plasmid replicon signatures of the isolates. We found that although ST15, ST14 and ST147 predominate, almost one-third of isolates came from STs considered infrequent in Portugal. We establish there are many AMR determinants, which are evolving over time, and our work provides a baseline set of variants for future monitoring and epidemiological studies in Portugal and wider Europe.

Results

In silico ST diversity and population structure of *K. pneumoniae* in Portugal.

This study includes a total of 509 Kp isolates. Of these, 459 are clinical isolates from hospitals in the southern (n = 378), central (n = 20) and northern (n = 61) regions of Portugal (Table 1). These were compared to isolates collected from veterinary clinics (n = 41) and environmental wastewater (n = 9) from the southern region, thereby broaden-ing insights into the incidence of Kp in Portugal. The isolates were collected between years 1980 and 2019, but the majority (n = 455, 89%) were collected between 2000 and 2019 (Table 1). Of the O antigen types, O1 was dominant with O1v1 (44%) and O1v2 (20%) followed by O2v2 (12%) and O2v1 (10%) serotypes (Table 1). Seventy-seven different STs were inferred, with the most frequent being ST15 (n = 161), ST147 (n = 36), ST13 (n = 26) and ST14 (n = 26) (Table 1). Globally significant clonal group 258 had little presence, with 4% of isolates belonging to ST11 and none to ST258 or ST512. A high proportion of isolates (31%) were from low frequency STs (each < 2.0%). Out of 2,926 possible unique ST pairs, only 0.7% differed by a single allele, whereas 22.7%, 32.8% and 25.7% differed by 4, 5 and 6 alleles, respectively (Fig. S2). O serotypes were linked strongly with STs (Fig. S3). Only 10 STs had isolates with different O serotypes. ST15 (n = 161) had 149, 9 and 3 isolates with O1v1, O1v2 and O1/O2v1 serotypes, respectively. ST13 (n = 26) had 25 and 1 isolates with O1v2 and O1v1 serotypes, respectively. ST14 (n = 26) had 24 and 2 isolates with O1v1 and O1/O2v1 serotypes, respectively. ST348 (n = 22) had 19 and 3 isolates with O1v1 and O1/O2v1 serotypes, respectively. ST11 (n = 19) had 13, 5 and 1 isolates with O2v2, O3b and O4 serotypes, respectively. There were a further five STs (34 isolates) containing isolates with different O serotypes. The predominant capsular K serotypes were KL112 (14%) and KL24 (19%), linked to the common ST15. When considering only the unique ST and K group combinations (n = 105), the most frequent K serotypes were KL30, KL10, and KL24 observed in 7, 7 and 5 STs, respectively.

Phylogenetic analysis.

The overall phylogenetic tree (n = 509) is largely congruent with ST type, but the propensity of Kp to recombine, means that the branch support values decline rapidly as one moves from leaves

Characteristic	N	%
Region		
Centre	20	3.9
South	428	84.1
North	61	12.0
Collection dates		
1980–1982	46	9.0
1990–1999	13	2.6
2000–2009	169	33.2
2010–2019	281	55.2
Sequence type (ST)		
ST15	161	31.6
ST147	36	7.1
ST13	26	5.1
ST14	26	5.1
ST348	22	4.3
ST307	20	3.9
ST11	19	3.7
ST231	16	3.1
ST70	16	3.1
ST45	12	2.4
Other**	155	30.5
Inferred O serotypes		
O1	329	64.6
O2	112	22.0
O3b	17	3.3
O4	13	2.6
O5	13	2.6
OL101	12	2.4
Unknown	10	2.0
O3/O3a	3	0.6
Inferred K serotypes		
KL24	96	16.7
KL112	71	14.8
KL64	35	7.1
KL62	32	6.4
KL3	26	5.6
Other	249	49.4
Carbapenemases and carbapenem-hydrolyzing beta-lactamases		
None	395	77.6
KPC-3	101	19.8
OXA-181	6	1.2
GES-5; KPC-3	5	1.0
OXA-48	1	0.2
GES-5	2	0.4
NDM-1	2	0.4
Aminoglycoside resistance genotypes		
AAC(3)-II; APH(3')-I; APH(6)-I;	154	30.3
AAC(6')-I; ANT(3")-I; APH(3')-I; APH(6)-I;	64	12.6
APH(3')-I; APH(6)-I;	33	6.5
AAC(3)-II; ANT(3")-I; APH(3')-I; APH(6)-I;	30	5.9
ANT(3")-I;	26	5.1
AAC(3)-II; AAC(6')-I; ANT(3")-I; APH(3')-I; APH(6)-I;	24	4.7
AAC(3)-II;	19	3.7
AAC(6')-I; ANT(3")-I;	18	3.5
AAC(3)-II; AAC(6')-I; ANT(3")-I;	10	2.0
ANT(3")-I; APH(3')-I; APH(6)-I;	9	1.8
Continued		

Characteristic	N	%
AAC(3)-II; ANT(3")-I;	8	1.6
AAC(3)-II; ANT(3")-I; APH(3')-I;	8	1.6
AAC(3)-II; APH(3')-I;	8	1.6

Table 1. Baseline characteristics for the 509 *K. pneumoniae* study isolates. **Includes ST1138 with 7 isolates.

to root (Fig. S3). Individual ST trees (Fig. 1A–D) identified several well supported geographic clades, including two previously reported[8,14] potential outbreaks in Lisbon and Vila Real hospitals involving representatives of ST147 and ST15. For the most abundant sequence type ST15 (n = 161), the tree shows multiple clades differentiated by K serotypes (Fig. 1A). Nearly all isolates carried *bla*$_{CTX-M-15,}$ except isolates from years 1980 to 1982 and 5 of 14 isolates from years 2007 to 2018 with *bla*$_{KPC-3}$. Due to the low number of isolates spanning years 1990 to 1999, our dataset may not fully reflect the beta-lactamase diversity during that period. It has been suggested that *bla*$_{TEM-10}$ was the dominant ESBL gene at that time, and most of our isolates (11/16) carried it. The smallest clade (n = 9) was from north Portugal and was distinguished by KL19 and O1v2 serotypes with five isolates carrying *bla*$_{KPC-3}$. The other *bla*$_{KPC-3}$ carrying ST15 isolates (n = 9) were phylogenetically very distant and part of a larger clade (n = 67) with K112 and O1v1 serotypes. Interestingly, these nine isolates carried both *bla*$_{KPC-3}$ and *bla*$_{CTX-M-15}$, in contrast to all other relevant isolates, which carry either *bla*$_{KPC-3}$ or *bla*$_{CTX- M-15}$ (Fig. 2). Four of the nine wastewater samples belonged to ST15, and were dispersed amongst clinical samples in the ST15 phylogenetic tree (Fig. 1A).

Most ST14 isolates (n = 20/26) fell within two dominant clades (Fig. 1B); the first clade (n = 11) was sourced from northern Portugal in year 2018 and contains the K16 locus, and the second (n = 9) was distinguished by the K2 serotype and spanned years 1980 to 2010. Only the second clade had *bla*$_{KPC-3}$ carrying isolates (n = 6). Unlike for other STs, the genes used in phylogenetic reconstruction of ST14 were relatively concentrated within a 3.60 Mbp to 3.75 Mbp chromosomal region based on the NC_016845.1 Kp assembly (Fig. S1).

All ST147 (n = 36) isolates had KL64 and O2v1 serotypes and belonged to two main clades. One clade consisted of northern region isolates (n = 15) from year 2018, all of which carried *bla*$_{KPC-3}$, but only one had *bla*$_{CTX-M-15}$ (Fig. 1C). All isolates in this clade carried the siderophore yersiniabactin (*ybt 16; ICEKp12*) locus related to virulence. A second clade (n = 20) was sourced from the southern region and spanned years 1995 to 2016. Only 11 isolates in this clade carried *bla*$_{KPC-3}$ and none had the yersiniabactin locus. Similarly, the ST13 tree had two clades characterised by the KL3 serotype (Fig. 1D). One clade (n = 10) was sourced from multiple hospitals in year 2019, and all isolates carried *bla*$_{KPC-3}$ and yersiniabactin loci. Another clade (n = 13) spanned years 1982 to 2013 and came from the southern region. Only two isolates in this clade carried *bla*$_{CTX-M-15}$, and none had other ESBL genes, *bla*$_{KPC-3}$, or the yersiniabactin locus. Unlike the remaining ST13 isolates, these two phylogenetically distant isolates had inferred K57 and K30 serotypes.

Our dataset included four isolates of type ST1138, an ST which was previously described at two hospitals in Portugal[13]. We also had previously undescribed ST6003 (n = 3), which differs from ST1138 by an *rpoB* allele. All seven isolates were collected in 2012 from the same hospital in Lisbon. Remarkably, one of the ST1138 isolates has no AMR genes or mutations apart from *bla*$_{SHV-36}$. This suggests its ancestors were newly introduced into the hospital system. This original strain has split into at least two lineages, both carrying *bla*$_{KPC-3}$, but it is unclear if this acquisition was independent in each lineage. There are at least two inferred lineages because the isolates from overlapping dates have one of two types of *rpoB* alleles (denoted as alleles 1 or 46). We did not identify any post-2012 ST1138 isolates in our data, but it seems to have been present in Portuguese isolates collected in year 2011[13]. Apart from ST6003, our dataset had two further one allele variants of ST1138: ST514 (n = 2) collected from hospital patients in the 1980s and ST323 (n = 1) collected from wastewater in 2018. Both ST514 and ST323 came from the same geographical area as sequence types ST1138 and ST6003.

Plasmid clustering. Of the 48 replicon families identified by PlasmidFinder software, six occurred in at least 10% of isolates (IncFIB 98%, IncFII 96%, IncFIA 55%, IncR 42%. IncHI1B 11%, IncN 10%). An average of 4 replicons (range: 0–9) were identified per isolate with only three isolates having none. To understand the pattern of plasmid replicons across isolates we applied UMAP dimensional reduction methods[29] to presence-absence matrices of replicon family (n = 48) and unique replicon sequences (n = 240). The results based on families and exact replicon sequences were consistent; so here we present the higher resolution analysis using replicon sequences (Data S1). We observed nine clear and robust clusters (replicon clusters, RCs) (Fig. 3A). RC prevalence varied over time (Fig. 3B) with RC9, the cluster with 94% *bla*$_{KPC-3}$ carriage, becoming dominant between years 2010 and 2019. By overlaying RC type onto the overall and ST specific phylogenetic trees, mosaic distributions were revealed, further supporting the movement of plasmids among Kp lineages (Fig. 1A–D).

RC1 contained 57% of all isolates (n = 288), but very few (n = 9) isolates had *bla*$_{KPC-3}$; although 59% carried ESBL encoding genes, which is consistent with early isolates dominating this cluster. The frequency of isolates in this cluster has declined substantially after year 2001. RC1 has a large diversity of plasmid replicons including IncFIB(K) (n = 214), IncFII(K) (n = 174), IncR (n = 108), and CoI(pHAD28) (n = 108) with each replicon family consisting of multiple distinct phylogenetic clades making RC1 interpretation complex. In contrast, the RC9 (n = 87) cluster has 25 different STs, including the dominant ST15, ST147 and ST14 types. Almost all Kp in RC9 cluster (94%, 82/87) carry *bla*$_{KPC-3}$. This gene is absent in 5 isolates: ST14 from 2007 and 2010; ST416 from 2011; ST1138 from 2012; and ST359 from 2018. The key signature of RC9 is the presence of variants of FIA(pBK30683)

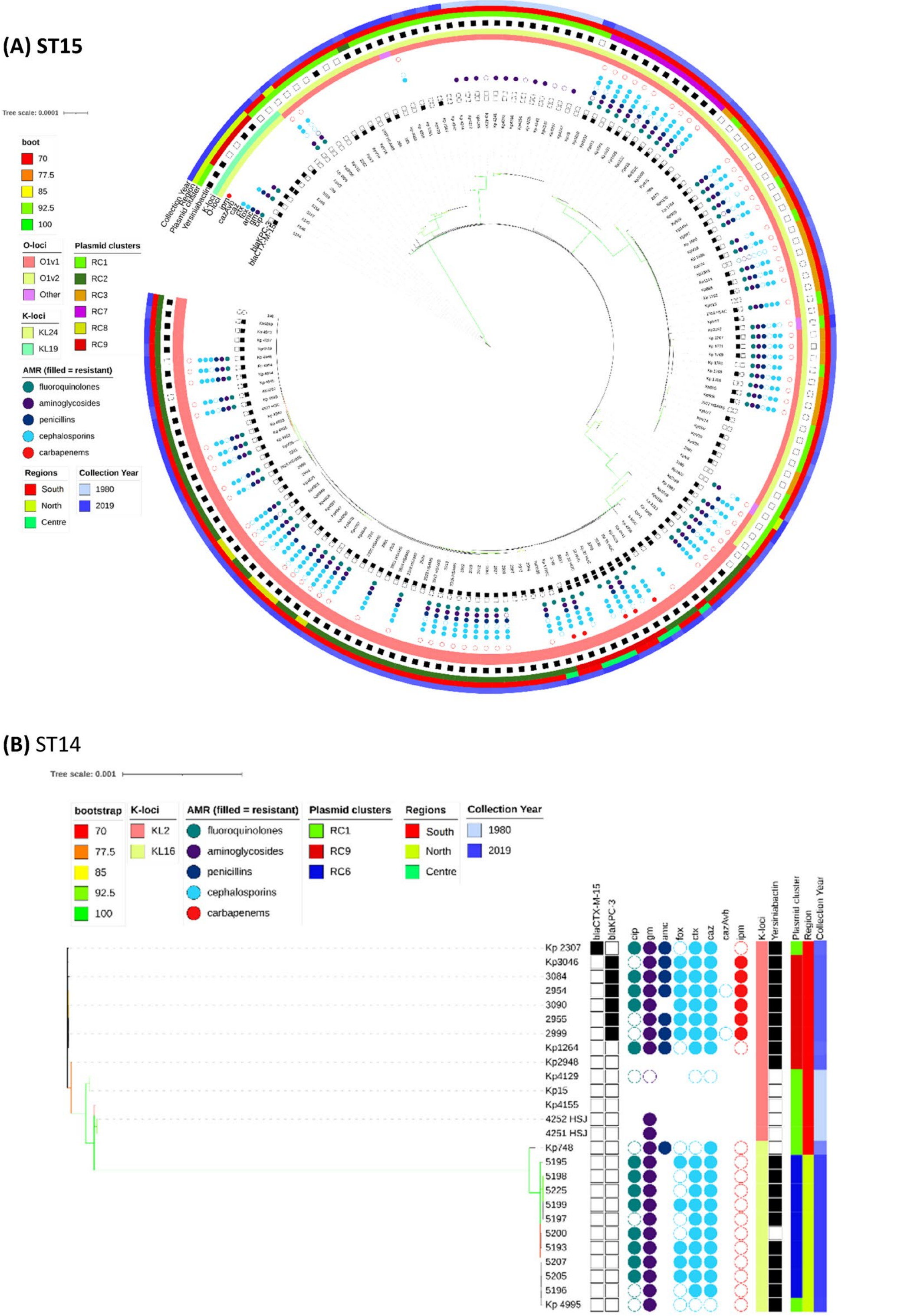


Figure 1. Phylogenetic trees of the most commons sequence types (STs), their antimicrobial resistance (AMR) phenotype, and carbapenemase and ESBL genotypic profiles. K-Loci refer to inferred K serotypes. Branch colours represent bootstrap support values.

and FII(pBK30683) replicons (Data S1). For isolates with $bla_{KPC\text{-}3}$ in RC9, more than a third (31/87) were diverse STs collected between years 2010 and 2019, but possessing neither *gyrA* mutations, nor an *aac(6')-Ib-cr* gene.

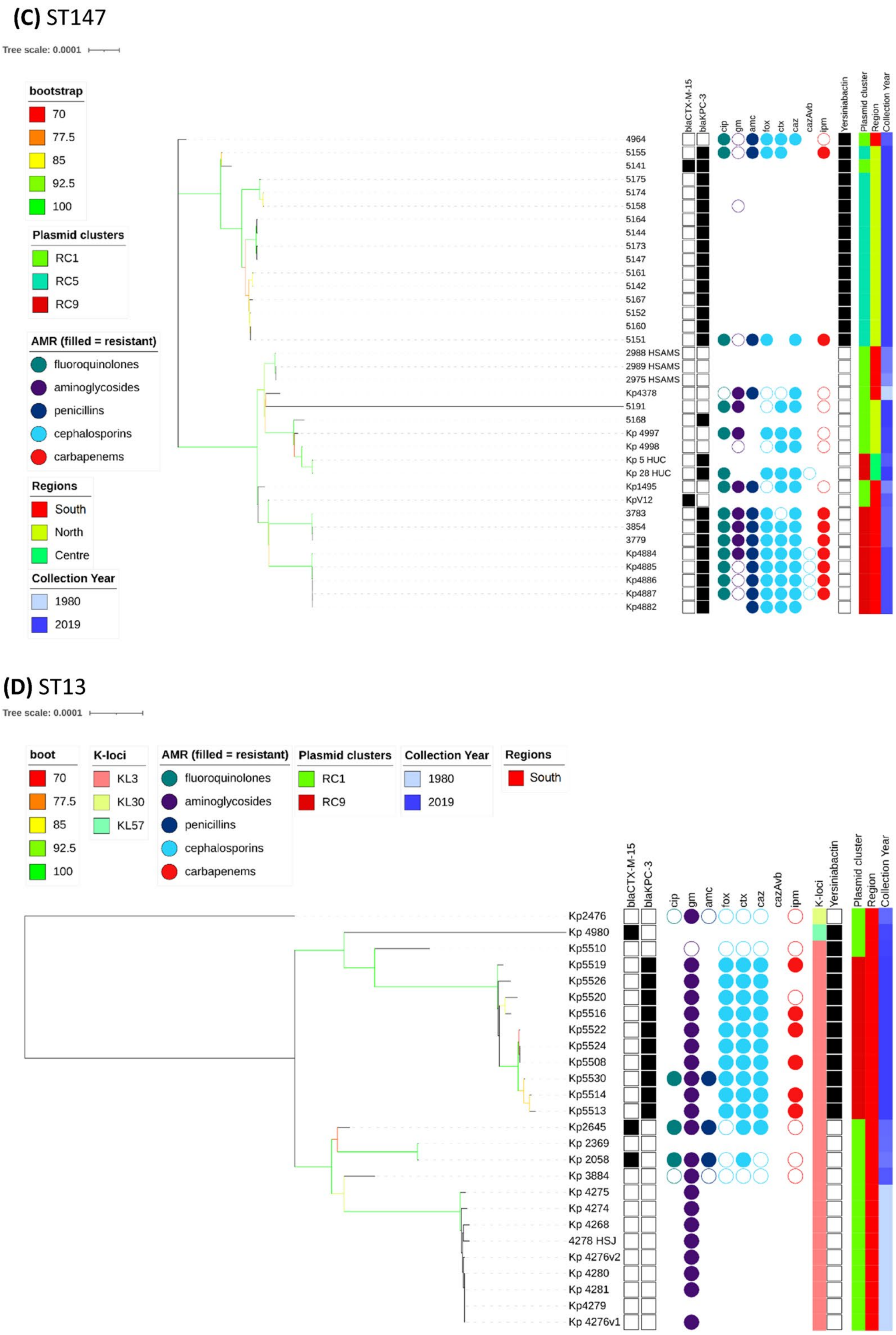


Figure 1. (continued)

These isolates were susceptible to fluoroquinolones with average inhibition zone diameters of 20 mm among 27 tested isolates.

The RC5 cluster (n = 14) consisted of ST147 carrying *bla*$_{KPC-3}$ and was congruent with the clade sourced from northern Portugal in year 2018 (n = 15). Of these northern Portuguese isolates, 14 belonged to RC5 and one belonged to RC1 (Fig. 1C). The RC5 cluster is identifiable by two variants of IncN and IncFIB(pKPHS1) replicons (Data S1). IncN is uncommon in our dataset (51/2222 replicons) and its strong link with *bla*$_{KPC-3}$ suggests, in line with earlier reports[15,30], that *bla*$_{KPC-3}$ may be mobilized by IncN plasmids. In contrast, other clusters had

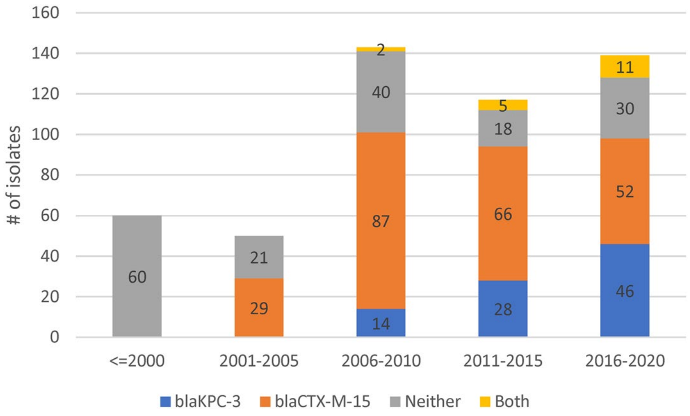


Figure 2. The most common ESBL ($bla_{CTX-M-15}$) and carbapenemase (bla_{KPC-3}) genes across the 509 Kp isolates by year group.

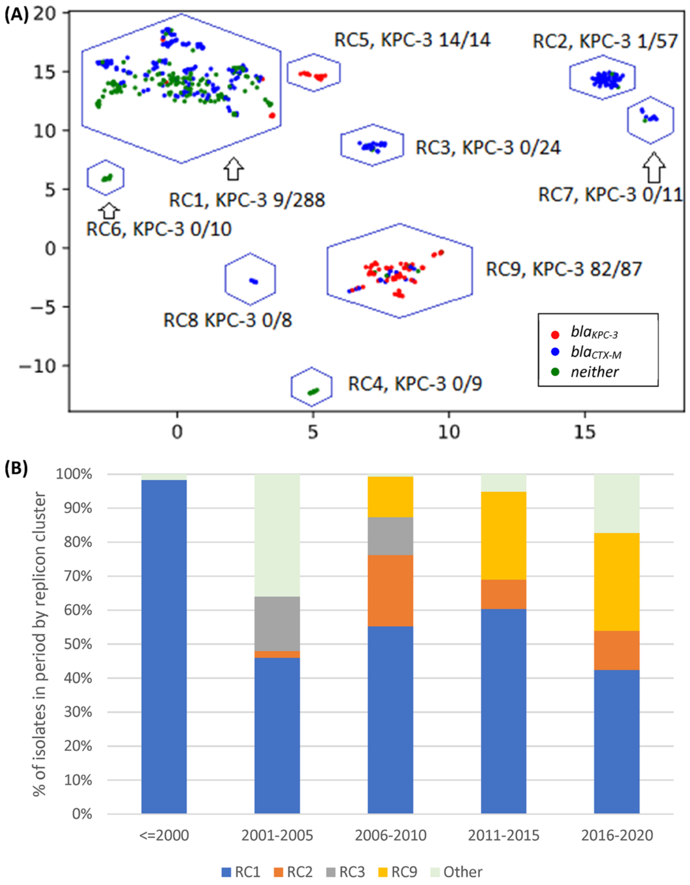


Figure 3. (**A**) Clustering of isolates by their plasmid replicon (replicon clusters, RC) and antimicrobial resistance (AMR) genotypic profiles, revealing differentiation by carriage of *blaKPC-3* and *blaCTX-M-15* genes. X and Y axis are dimensions on which full data is projected, they are unitless; (**B**) Abundance of isolates from different plasmid clusters.

homogenous ST types: RC2, RC3, RC7 and RC8 consist of 57, 24, 11, and 8 ST15 isolates; RC6 consists of ten ST14 isolates; and RC4 of nine ST12 isolates. RC3 is interesting because despite having isolates from years 2003 to 2014, only one (1/24) had a bla_{SHV} type beta-lactamase (bla_{SHV-11}). These genes are very common (87.4%) and are normally chromosomal, which suggests either the loss of bla_{SHV} or simultaneous circulation of several strains.

All RC2 isolates carried IncFIB(K) and IncFII(pKP91) replicons, where the associated variants were different to those from the RC1 replicons. The RC2 IncFIB(K) variant was also present in RC8 and RC7 isolates. RC7 isolates also carried a distinct variant of a Col440I replicon, and a variant of IncR is shared with RC1, RC2, RC3 and RC8. RC8 carried a distinct version of IncFII(K) and IncFIB(pQil) replicons. Finally, in addition to shared IncR and ColpVC variants, RC3 also carried IncHI1A and IncHI1B(R27) replicons, which were absent in other clusters (Data S1).

Antimicrobial resistance genotypes and phenotypes. *In vitro* AMR profiles and genotypes were analysed for the 509 Kp isolates, with some gaps depending on the decade of phenotypic assessment (Table S1). Most isolates underwent antibiotic susceptibility testing for aminoglycosides (n = 356, 68.1%), cephalosporins (n = 366, 71.2%), fluoroquinolones (n = 344, 66.9%), carbapenems (n = 296, 57.6%), and penicillins (n = 311, 60.5%) (Table S1). However, 121 isolates had no AMR susceptibility data. Since AMR testing was performed over multiple years, the concentration of active compounds in disks used might vary between isolates; all breakpoints were determined based on EUCAST v11.0 (2021)[31].

Beta-lactams. The majority (75%) of 304 tested isolates showed susceptibility to imipenem; an antibiotic used widely in hospital clinical practice in Portugal (Fig. 4). The resistance driver was likely bla_{KPC-3} (n = 106), which was carried on a Tn4401d transposon in nearly all cases (n = 101/106). In four isolates [ST15 (Kp5149), ST147 (Kp5147), ST34 (Kp5148) and ST461 (Kp5162)], the bla_{KPC-3} gene has undergone an identical inversion within the Tn4401d structure. Only one of these four isolates (ST147) was tested for imipenem resistance and was determined to be resistant (inhibition zone diameter of 6 mm). This observation suggests the inversion did not significantly impair the resistance conferred by the presence of bla_{KPC-3}. Additionally, our dataset contained three isolates without carbapenemase genes but resistant to imipenem. These three isolates had no clear commonality between them, nor clear distinction from susceptible isolates with a similar genotype.

Resistance to cephalosporins was widespread with 68% of 335 tested isolates showing resistance to cefotaxime, a third-generation cephalosporin, driven likely by the presence of class A ESBL $bla_{CTX-M-15}$ (n = 213, 41%) (Fig. 4). Ceftazidime had an even higher resistance prevalence of 91%. This drug was used widely in the 1990's for Pseudomonas outbreaks, but is currently rarely used in monotherapies. The combination of ceftazidime and carbapenemase inhibitor avibactam (cazAvb) was rarely resistant (1/46 tests). The resistant isolate (ST15) carried bla_{OXA-1}, bla_{TEM-1}, bla_{SHV-28} and had truncated porin gene *ompK36*. Of the imipenem resistant isolates, twelve were tested for cazAvb and all were susceptible.

Finally, the second-generation cephalosporin cefoxitin presents an interesting case because its resistance profile was frequently inconsistent with its genomic or genotypic profile (Fig. 4). Of 366 tested isolates, 8 carried AmpC beta-lactamase bla_{DHA-1}, which is associated with strong inhibition of cefoxitin. Consistent with this, these 8 isolates had a cefoxitin inhibition zone diameter of 6 mm. However, the inhibition zone diameter of the tested bla_{KPC-3} carrying isolates (n = 74) was on average 13 mm, which is below the 19 mm breakpoint, but well above the total inhibition diameter of 6 mm. In isolates with class A (non-broad-spectrum, broad-spectrum, ESBL) and D beta-lactamases, the susceptibility to cefoxitin varied vastly with isolates containing all four categories having almost uniform distribution of inhibition zone diameters between 25 and 6 mm. We were able to trace some of this inconsistency to 83 isolates with bla_{SHV-28}. The mean diameter for isolates with class A broad spectrum, class A ESBL and class D genes is 22 mm. If this genotype also included bla_{SHV-28} (class A), the mean inhibition zone diameter reduced to 16 mm, but this effect is absent for other bla_{SHV} variants.

Fluoroquinolones. Of the 307 isolates tested for ciprofloxacin susceptibility, 288 (94%) were resistant (Fig. S4). Five isolates did not have any common fluoroquinolone resistance determinants and were all susceptible with between 22 and 30 mm inhibition zone diameters. By far the strongest determinant of resistance was the presence of mutations in type II topoisomerase *gyrA*. *In silico* screening with Kleborate software identified simultaneous mutations GyrA83F and GyrA87A (n = 150), and single mutations GyrA83I (n = 91) and GyrA83Y (n = 15) as the most common ones. Nearly all isolates with *gyrA* mutations had inhibition zone diameters below 10 mm, whereas the resistance breakpoint is 22 mm. In 54 tested isolates that had both the *gyrA* mutation and *qnrB* gene, nearly all had inhibition zone diameters of 6 mm, demonstrating clear compounding of resistance. The *oqxAB* genotype gives a small decrease in susceptibility, but most of such isolates were still susceptible. Finally, *qnrB* and *qnrS* each substantially decreased inhibition zone diameter. In our dataset these genes were present only in combination with *oqxAB,* leading to reduced inhibition zone diameters of ~ 50% compared to *oqxAB* alone. As we did not have isolates with both *qnr* genes, we could not confirm if their effect is cumula-tive. Further, 36 and 24 isolates were tested with levofloxacin and norfloxacin, respectively, with 42% of each set showing susceptibility. In contrast to ciprofloxacin, levofloxacin was only moderately affected by *gyrA* mutations with an average inhibition zone diameter reduced to 14 mm (Fig. S4).

We examined the diversity of the *parC* gene which has been associated with fluoroquinolone resistance [5,32], and found 244 isolates carrying both the ParC80I mutation and *gyrA* mutations. The *parC* alleles were ST specific except for three ST307 outlying isolates that differ by a single SNP from the other seventeen ST307 and fifteen ST15 isolates (Fig. S5) which suggests a lack of selective pressure on the gene. A tree constructed using *gyrA* sequences formed clear clades for the abundant ST14 and ST15 types (Fig. S5). ST14 isolates formed two major clades, one linked to decades 1980's and 2010's (n = 14) and another from the 2000's (n = 15). ST15 formed three clades, one linked to the 1980s (n = 13), and two others post-year-2000 (n = 18, n = 131). There was additional evidence of selective pressure in other STs for which we had fewer isolates (Fig. S5). We could not test the effect of *parC* mutations on resistance because all isolates with mutations in this gene also had mutations in *gyrA*. Finally, 237 isolates (46%) had an *aac(6')-Ib-cr* gene, of which 132 were tested for ciprofloxacin resistance and 128 were resistant. This gene had moderate impact on disk diameter, but this impact is sufficient for isolates to fall below resistance threshold. While the majority of *aac(6')-Ib-cr* carrying isolates also had *gyrA* mutations,

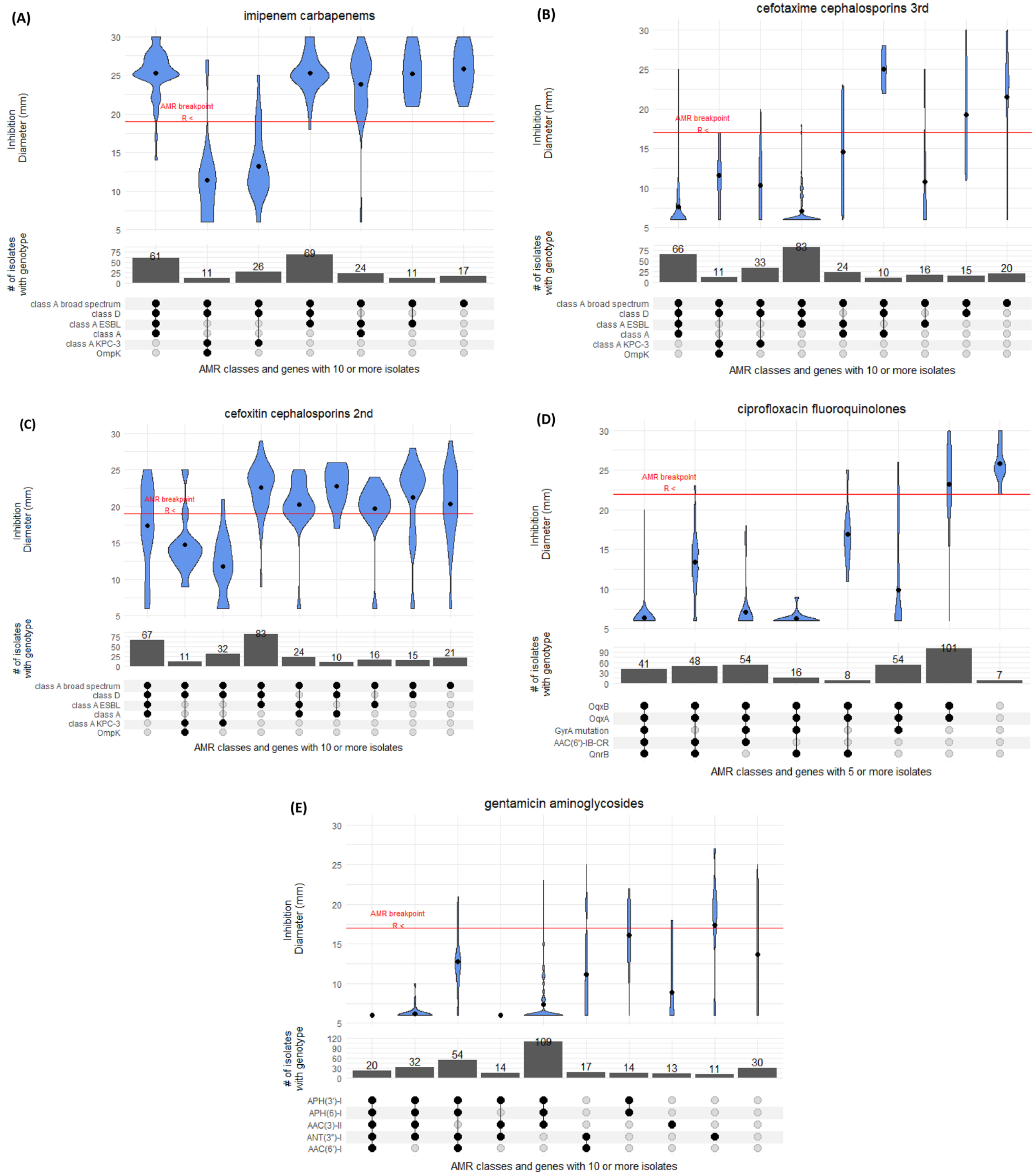


Figure 4. Distribution of inhibition zone diameters for different genotypes for (**A**) imipenem, (**B**) cefotaxime, (**C**) cefoxitin, (**D**) ciproflocaxin, and (**E**) gentamicin antimicrobials.

we found 48 isolates with an *oqxAB*/*qnrB*/*aac(6')-Ib-cr* genotype for which mean inhibition zone diameter was 13 mm compared to 17 mm for 8 isolates with an *oqxAB*/*qnrB* genotype.

Tetracycline. We tested Kp isolates for tetracycline (n = 57) and tigecycline (n = 48) resistance. While breakpoints for tetracycline are not standardised[31], the results indicate very high resistance to the antibiotic (Fig. S4). The only five isolates susceptible to tetracycline were collected between years 1980 and 1982, which suggests in the remaining 52 isolates that resistance is acquired rather than intrinsic. Isolates with *tetABD* efflux pumps (n = 31/57) had marginally more resistant profiles. In contrast to tetracycline, most of tigecycline tests inhibition zone diameters were located just below the breakpoint, with only a few isolates with < 10 mm (Fig. S4). Kp does not have an EUCAST zone diameter breakpoint, so we used the *E. coli* 18 mm breakpoint instead. Interestingly, the *tetABD* efflux pumps did not reduce the inhibition zone diameter. Isolates without these efflux pumps were marginally more (mean of 14 mm versus 17 mm), not less, resistant.

Aminoglycosides. Among isolates tested with gentamicin (n = 349) and amikacin (n = 29), 89% and 55% were resistant, respectively. In those isolates that had no known gentamicin resistance determinants (n = 28), the average inhibition zone diameter was 15 mm versus an established breakpoint of 18 mm (Fig. 4). Two acetyltransferases were present in our isolates ((AAC(6')-I, n = 283; AAC(3)-II, n = 291), of which AAC(3)-II was a stronger determinant of resistance, with nearly all inhibition zone diameters below 10 mm. AAC(6')-I reduced the inhibition zone diameter from 15 to 12 mm, and the effect was cumulative with AAC(3)-II. Nucleotidyltransferase ANT(3")-I and phosphatases APH(6')-I and APH(3')-I, even combined, appear to have a marginal effect on resistance. The 16S rRNA methyltransferase gene *armA* was present in two isolates, and the gene *rmtB* was present in two additional isolates.

Virulence determinants. Using VFDB, Kleborate and BLAST tools, we examined the dataset for virulence genes[33–35]. The type 1 fimbriae locus, *fimABCDH*, was present in nearly all isolates (501/509, 98%). It was absent in isolates from ST960 (4/4), ST15 (3/161) and ST76 (1/3). Similarly, the type 3 fimbriae locus, *mrkABCDF*, was present in all isolates except for one (ST147). In contrast, iron uptake locus *kfu* was present in 245 (50%) isolates. This locus was perfectly correlated with ST; no ST had simultaneously *kfu* positive and negative isolates. Among the most frequent STs, ST15 (n = 161), ST13 (n = 26) and ST14 (n = 26) isolates had a *kfu* locus, but it was absent in ST147 (n = 36), ST348 (n = 22), and ST307 (n = 20) isolates. There were only a few uncommon virulence genes. For example, we did not find any *rmpA* and *rmpA2* regulators of hypermucoviscosity. Two isolates (Kp4248 and KpV9) had aerobactin siderophores: one with iucA2 (ST3) and the other with iuc3 (ST3027). Both isolates had limited known AMR determinants. ST3 had only bla_{SHV-1} and ST3027 had APH(6')-Ia/d, *tet(A)* and bla_{SHV-33}. The salmochelin locus *iroBCDEN* was found only in some ST48 isolates (n = 7/9), all of which carried ESBL $bla_{CTX-M-15}$. Of these 7 isolates, 6 came from the same Lisbon hospital between years 2005 and 2009. Yersiniabactin was present in 56% of isolates, and was carried most frequently on integrative conjugative elements ICEKp3, ICEKp4 and ICEKp12 at 21%, 15% and 6% of all isolates, respectively. Additional screening against the VFDB virulence database only revealed *astA* and *cseA* genes in one isolate.

Discussion

In this work we have analysed whole genome sequence data from 509 Kp isolates collected between years 1980 and 2019 from the hospital systems across southern, central, and northern regions of Portugal, as well as from veterinary clinics and a sewage treatment plant in the southern region. Because this is one of the largest in-country collections sequenced, it allowed us to investigate and understand the temporal and spatial genetic diversity of this important pathogen. We observed that 31% of isolates belonged to STs that are considered infrequent in Portugal. If the ST diversity of the dataset was driven by mutations within ST determining genes, we would expect that most samples differ by a single allele, but this was not the case. This observation is further supported by SNP distances between different STs. Instead, the diversity is more likely to be driven by either recombination or coexistence of many strains. The simultaneous presence of so many STs in the country is suggestive of importation and large environmental or human reservoirs of infection. The latter is consistent with high rates of colonisation observed in different countries and settings[36]. The non-human sourced isolates from wastewater and animal settings did not standout in the analysis, which suggests a flow of Kp between humans and environmental reservoirs. However, the limited number of non-human sourced isolates did not allow us to determine the direction of this flow.

Our WGS sequencing analysis revealed insights into AMR genotyping and phenotyping. During the period with best isolate coverage, years 2000 to 2019, the dominant beta-lactam and carbapenem resistance determinants were $bla_{CTX-M-15}$ (41%) and bla_{KPC-3} (21%). In isolates from years 1990 to 1999, a majority (11 of 16) had bla_{TEM-10}, which is thought to be the dominant ESBL during that period. Our analysis of plasmids reveals a complex and mosaic distribution across isolates suggestive of active selection between them. The increase in prevalence of bla_{KPC-3} has been accompanied by a decrease in the frequency of older $bla_{CTX-M-15}$, which encodes narrower spectrum beta-lactamase. We observed very strong clustering of isolates by their detected plasmid replicons. Plasmid naming nomenclature is based on shared replication mechanisms and incompatibility[37], so we expected to observe some structure. The replicons themselves revealed a rigid pattern that is clinically relevant due to carriage of bla_{KPC-3} on two types of plasmids. Isolates with FIA(pBK30683) and FII(pBK30683) replicons were possible sources of bla_{KPC-3}, while isolates with IncN and IncFIB(pKPHS1) were restricted to ST147 types. However, the isolates with bla_{KPC-3} had an exact same replicon allele present for at least 10 years and formed a very clear cluster of isolates (denoted as replicon cluster 9).

While our AMR test results may suffer from changes in the amount of active compound in test disk assays across the years, we did find interesting and robust results. As expected, we found that bla_{KPC-3} was the dominant carbapenemase gene[8,12,13], but we also observed a reduction in carriage of ESBL bla_{CTX-M} in those isolates that acquired $bla_{KPC-3.}$ This displacement of bla_{CTX-M} indicates active selection, and co-carriage of bla_{CTX-M} and bla_{KPC-3} was very rare. While AMR genotype was largely consistent with phenotype, we observed that cefoxitin, a retired second-generation cephalosporin shows moderate activity in isolates lacking class C beta-lactamases. We also found that nearly half of the isolates in replicon cluster 9, which had almost universal carriage of bla_{KPC-3}, were susceptible to fluoroquinolones as they lacked *gyrA* mutations or other resistance factors. More generally, such insights may offer opportunities for additional treatment of infections with bla_{KPC-3} carrying Kp.

Overall, our work has provided temporal and spatial insights into Kp STs and AMR related genes and plasmids circulating in Portugal. We found a large diversity of STs, with ST15 and ST147 being the most frequent (< 40%), but almost one-third of isolates had uncommon types (< 2% frequency). Dominant beta-lactamase genes are changing over time due to changes in drug utilisation and plasmid changes. The bla_{OXA-9} and bla_{TEM-1} of the 1980s were displaced by $bla_{CTX-M-15}$ in 2000's which in turn were replaced by bla_{KPC-3}. These insights reinforce the need for genomic sequencing and tools to assist surveillance and clinical decision making.

Methods

Isolate collection, library preparation and sequencing. The isolates (n = 509) were identified between years 1980 and 2019 from 16 hospitals in Lisbon and its metropolitan area (Southern region), Coimbra (Central region), and Porto and Vila Real (Northern Portugal), except for 9 isolates from Beirolas wastewater (Lisbon) and 41 samples from veterinary clinics (Lisbon) (Table 1). The isolates were cultured as described previously[8]. The set of isolates represent a convenience sample accumulated over 40 years, with the collection site known for the majority (74%). Isolates with known collection site were sourced from blood (32%), urine (31%), rectal screening swabs (10%), pus (7%) and wastewater (7%). Clinical isolates obtained from hospitals were identified at local clinical microbiology laboratories and sent to the Faculty of Pharmacy (University of Lisbon; FFUL) for further phenotypic and genotypic analysis. Given the wide temporal span of the isolates, the initial identification methods that were employed at local laboratories vary across isolates, but were based on biochemical identification methods (e.g., API, Vitek). DNA was extracted from strain cultures grown overnight at 37ºC on Mueller–Hinton Agar. DNA extraction was carried out using the Cetyl trimethylammonium bromide method[38]. Library preparation of the DNA samples was performed using a QIAseq FX DNA library kit, following the manufacturer's protocol. WGS was performed on Illumina HiSeq (paired end 150 bp) through The Applied Genomics Centre (London School of Hygiene and Tropical Medicine)[39]. Only those isolates which Kleborate software (v 2.1.0)[40] identified as Kp were used for further analysis.

Antimicrobial susceptibility testing. Antimicrobial susceptibility testing at FFUL was carried out using the Kirby-Bauer disk diffusion method as per the European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines for performance and interpretation of antimicrobial susceptibility testing (v11.0, 2021)[31]. AMR testing was performed over multiple years, and therefore the concentration of active compounds in disks used might vary between isolates. The AMR testing was not performed for 121 isolates, and not all isolates were tested for the same antimicrobials. The Pearson correlation coefficient was used to assess correlations between disk inhibition zone diameters for different antimicrobials.

Genome assembly, annotation, and genotyping. Raw Illumina reads were assembled using Unicycler software (v0.4.8)[41], with assembly fragmentation and completeness assessed against 440 core genes of enterobacterales (enterobacterales_odb9) using Busco software (v4)[42]. The assessed quality of assemblies was high (median N50: 284Kbp) with all but one having complete single copies of > 97% genes in the Busco reference set. Assemblies were annotated using Prokka software (v 1.14.6)[43], combined with the *Klebsiella* specific reference genes set[44]. O and K antigen serotypes, genomic AMR, virulence (e.g., ICEKps), and sequence type (ST) profiles were analysed and inferred *in silico* using Kleborate[40] and AMRFinder (v3.8.4)[45] software with associ-ated databases (accessed October 2020). We have also used Abricate software (v1.0.1)[46] with the virulence factor database VFDB (accessed March 2021) to find additional virulence genes. Plasmid detection and classification was performed using Plasmidfinder software (v2.1.1)[43].

Phylogenetic analysis. The recombination and horizontal gene transfer within Kp can complicate phylogenetic analysis, and a widely used Kp MLST scheme[22] does not allow for a high-resolution phylogeny. We have observed that regions linked to transposons and other genes relating to mobile genetic elements tended to produce high number of SNPs, which would bias phylogenetic reconstruction. For this reason, instead of performing phylogenetic reconstruction using SNPs called against a reference genome, we used a reference-free method. We focused on a subset of core genes defined by two conditions. First, in all isolates, the gene coding sequence length is within 1% of median gene length across isolates. Second, the genes have pairwise identity of > 99%. Based on these criteria, out of ~ 5,000 genes in Kp isolates, 1424, 2212, 2802, and 3170 genes were present in 100%, 99%, 95% and 90% of all 509 isolates. We used Shannon's entropy to identify a 100 (from the 1424) genes with the most diverse nucleotide sequences. These 100 genes were aligned used MAFFT (v7.467)[48] and the resulting alignments used to construct phylogenetic trees for the entire isolate set (n = 509). The trees were reconstructed using IQTREE (v2.0.3)[49] with 1000 bootstrap replicates for each tree. The scripts can be found at https://github.com/AntonS-bio/entropy. For each ST, the gene selection was performed separately. The location of each chromosomal gene on the Kp reference genome (NC_016845.1) (Fig. S1) was generated with BRIG software (v 0.95)[50]. Phylogenetic trees were visualised in ITOL and are available (https://itol.embl.de/shared/Zp28yLE9IuWB).

Statistical analysis. Statistical analysis was performed using R software (v4.0.3)[51]. Additional analysis was performed in Python (v3.6). For detection of replicon clusters, we created a presence/absence matrix with one row per isolate and one column per each unique replicon sequence. We performed dimensional reduction on this matrix using the Uniform Manifold Approximation and Projection (UMAP) algorithm[29] implemented in the R uwot package. We used both hamming and jaccard distance measures and a broad range of parameters to establish the robustness of our results. The cluster detection was performed using the DBSCAN algorithm implemented in R[52]. Analysis scripts are available on https://github.com/AntonS-bio.

Consent for publication. All authors have consented to the publication of this manuscript.

Data availability

All sequencing data is available from European Nucleotide Archive project PRJEB47288. All AMR testing data is available in PATRIC database. Analysis scripts are available at https://github.com/AntonS-bio. High resolution phylogenetic trees are available at https://itol.embl.de/shared/Zp28yLE9IuWB.

Received: 1 November 2021; Accepted: 3 August 2022
Published online: 13 August 2022

References

1. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *K. Pneumoniae*, an urgent threat to public health. *Proc. Natl. Acad. Sci. USA.* **112**(27), E3574–E3581 (2015).
2. Wyres, K. L. & Holt, K. E. *K. Pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Curr. Opin. Microbiol.* **45**, 131–139 (2018).
3. Ramirez, M. S. & Tolmasky, M. E. Aminoglycoside modifying enzymes. *Drug Res. Updates.* **13**(6), 151–171 (2010).
4. Redgrave, L.S., Sutton, S.B., Webber, M.A., & Piddock, L.J.V. *Fluoroquinolone Resistance: Mechanisms, Impact on Bacteria, and Role in Evolutionary Success*, Vol. 22, Trends in Microbiology 438–45 (Elsevier Ltd, 2014). https://doi.org/10.1016/j.tim.2014.04.007.
5. Hooper, D. C. & Jacoby, G. A. Topoisomerase inhibitors: Fluoroquinolone mechanisms of action and resistance. *Cold Spring Harbor Perspect. Med.* **6**(9), a025320 (2016).
6. European Centre for Disease Prevention and Control. Antimicrobial resistance in the EU/EEA (EARS-Net)-Annual Epidemiological Report for 2019 [Internet]. 2020. https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2019.
7. Ferrari, C. *et al.* Multiple *K. Pneumoniae* KPC clones contribute to an extended hospital outbreak. *Front. Microbiol.* **10**, 556 (2019).
8. Perdigão, J. *et al.* Whole-genome sequencing resolves a polyclonal outbreak by extended-spectrum beta-lactam and carbapenem-resistant *K. Pneumoniae* in a Portuguese tertiary-care hospital. *Microbial. Genom.* **6**, 8896 (2020).
9. Logan, L. K. & Weinstein, R. A. The epidemiology of carbapenem-resistant enterobacteriaceae: The impact and evolution of a global menace. *J. Infect. Dis.* **215**(Suppl 1), S28-36 (2017).
10. Munoz-Price, L.S., Poirel, L., Bonomo, R.A., Schwaber, M.J., Daikos, G.L., Cormican, M. *et al. Clinical epidemiology of the global expansion of K. Pneumoniae carbapenemases*, Vol. 13, The Lancet Infectious Diseases 785–96 (NIH Public Access, 2013).
11. Aires-De-Sousa, M. *et al.* Epidemiology of carbapenemase-producing *K. Pneumoniae* in a hospital, Portugal. *Emerg. Infect. Dis.* **25**(9), 1632–1638 (2019).
12. Guerra, A. M. *et al.* Multiplicity of carbapenemase-producers three years after a kpc-3-producing k. Pneumoniae st147-k64 hospital outbreak. *Antibiotics* **9**(11), 1–11 (2020).
13. Manageiro, V. *et al.* Predominance of KPC-3 in a survey for carbapenemase-producing Enterobacteriaceae in Portugal. *Antimicrob. Agents Chemother.* **59**(6), 3588–3592 (2015).
14. Perdigão, J. *et al.* Genomic epidemiology of carbapenemase producing *K. Pneumoniae* strains at a northern portuguese hospital enables the detection of a misidentified klebsiella variicola kpc-3 producing strain. *Microorganisms.* **8**(12), 1–18 (2020).
15. Rodrigues, C. *et al.* KPC-3-producing *K. Pneumoniae* in Portugal linked to previously circulating non-CG258 lineages and uncommon genetic platforms (Tn4401d-IncFIA and Tn4401d-IncN). *Front. Microbiol.* **7**, 665 (2016).
16. Pires, D. *et al.* Evolving epidemiology of carbapenemase-producing Enterobacteriaceae in Portugal: 2012 retrospective cohort at a tertiary hospital in Lisbon. *J. Hosp. Infect.* **92**(1), 82–85 (2016).
17. Caneiras, C., Lito, L., Melo-Cristino, J. & Duarte, A. Community-and hospital-acquired *K. Pneumoniae* urinary tract infections in Portugal: Virulence and antibiotic resistance. *Microorganisms.* **7**(5), 83356 (2019).
18. Phelan, J. *et al.* The variability and reproducibility of whole genome sequencing technology for detecting resistance to anti-tuberculous drugs. *Genome Med.* **8**(1), 2258 (2016).
19. Phelan, J. E. *et al.* Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. *Genome Med.* **11**(1), 882 (2019).
20. Napier, G. *et al.* Robust barcoding and identification of Mycobacterium tuberculosis lineages for epidemiological and clinical studies. *Genome Med.* **12**(1), 114 (2020).
21. Guerra-Assunção, J. A. *et al.* Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife.* **2**(4), 558 (2015).
22. Diancourt, L., Passet, V., Verhoef, J., Grimont, P. A. D. & Brisse, S. Multilocus sequence typing of *K. Pneumoniae* nosocomial isolates. *J. Clin. Microbiol.* **43**(8), 4178–4182 (2005).
23. Wyres, K. L. *et al.* Identification of Klebsiella capsule synthesis loci from whole genome data. *Microbial Genom.* **2**(12), e000102 (2016).
24. Follador, R. *et al.* The diversity of *K. Pneumoniae* surface polysaccharides. *Microbial Genom.* **2**(8), e000073 (2016).
25. Bialek-Davenet, S. *et al.* Genomic definition of hypervirulent and multidrug-resistant *K. Pneumoniae* clonal groups. *Emerg. Infect. Dis.* **20**(11), 1812–1820 (2014).
26. Brisse, S. *et al.* Virulent clones of *K. Pneumoniae*: Identification and evolutionary scenario based on genomic and phenotypic characterization. *PLoS ONE* **4**(3), 2286 (2009).
27. Wyres, K. L. *et al.* Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *K. Pneumoniae*. *PLoS Genet.* **15**(4), e1008114 (2019).
28. Comandatore, F. *et al.* Gene composition as a potential barrier to large recombinations in the bacterial pathogen *K. Pneumoniae*. *Genome Biol. Evol.* **11**(11), 3240–51 (2019).
29. McInnes, L., Healy, J., & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 2018; http://arxiv.org/abs/1802.03426
30. Lopes, E. *et al.* Epidemiology of carbapenemase-producing *K. Pneumoniae* in northern Portugal: Predominance of KPC-2 and OXA-48. *J. Global Antimicrob. Res.* **22**, 349–353 (2020).
31. EUCAST: Clinical breakpoints and dosing of antibiotics-v 11.0. 2021. https://www.eucast.org/clinical_breakpoints/.
32. Geetha, P. V., Aishwarya, K. V. L., Mariappan, S. & Sekar, U. Fluoroquinolone resistance in clinical isolates of *K. Pneumonia* e. *J. Lab. Phys.* **12**(2), 121 (2020).
33. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: Hierarchical and refined dataset for big data analysis–10 years on. *Nucleic Acids Res.* **44**(D1), D694–D697 (2016).
34. Agarwala, R., Barrett, T., Beck, J., Benson, D.A., Bollin, C., Bolton, E. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2018.
35. Lam, M. M. C. *et al.* A genomic surveillance framework and genotyping tool for *K. Pneumoniae* and its related species complex. *Nat. Commun.* **12**(1), 1–16 (2021).

36. Huynh, B. T. *et al. K. Pneumoniae* carriage in low-income countries: Antimicrobial resistance, genomic diversity and risk factors. *Gut Microbes.* **11**(5), 1287–1299 (2020).
37. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. *Mobile Genetic Elements Associated with Antimicrobial Resistance* Vol. 31 (American Society for Microbiology, 2018).
38. Parish, T., Stoker, N.G., van Soolingen, D., de Haas, P.E.W., & Kremer, K. Restriction Fragment Length Polymorphism Typing of Mycobacteria. In *Mycobacterium Tuberculosis Protocols* 165–203 (Humana Press, 2003).
39. Spadar, A. *et al.* Methylation analysis of *K. Pneumoniae* from Portuguese hospitals. *Sci. Rep.* **11**(1), 6491 (2021).
40. Lam, M. M. C., Wick, R. R., Wyres, K. L. & Holt, K. E. Genomic surveillance framework and global population structure for *K. pneumoniae. bioRxiv.* **2**, 202012 (2020).
41. 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**(6), 197–215 (2018).
42. Seppey, M., Manni, M., & Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness. In *Methods in Molecular Biology* 227–45 (Humana Press Inc., 2019).
43. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**(14), 2068–2069 (2014).
44. Ehrlich, R. Prokka database maker. 2019. https://github.com/rehrlich/prokka_database_maker.
45. Feldgarden, M. *et al.* Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**(11), 550 (2019).
46. Seemann, T. Abricate: Mass screening of contigs for antimicrobial and virulence genes. https://github.com/tseemann/abricate.
47. Carattoli, A. *et al. In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**(7), 3895–3903 (2014).
48. Katoh, K. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**(14), 3059–3066 (2002).
49. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**(1), 268–274 (2015).
50. Alikhan, N.-F., Petty, N. K., Zakour, N. L. B. & Beatson, S. A. BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genom.* **12**(1), 1–10 (2011).
51. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.
52. Hahsler, M., Piekenbrock, M. & Doran, D. dbscan: Fast density-based clustering with *R. J. Stat. Softw.* **91**(1), 71158 (2019).

Acknowledgements

SC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1) and BBSRC UK (BB/R013063/1). TGC is supported by the Medical Research Council UK (Grant no. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC (BB/R013063/1). JPe is supported by Fundação para a Ciência e Tecnologia (ref. CEECIND/00394/2017 and UID/DTP/04138/2019). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

A.D., S.C., J.Pe. and T.G.C. designed the study. A.M., R.E., J.M.-C., G.L.S., C.P., C.M., M.P., and M.J.S. collected and processed isolates, including performing A.M.R. testing. S.C. and T.G.C. generated the sequence data. A.S. analysed the data, under the supervision of J.Ph., J.Pe., and T.G.C. A.S. wrote the first draft of the manuscript, with contributions from A.D., J.Pe. and T.G.C. All authors have edited manuscript drafts and agreed on the contents of the final version.

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/s41598-022-17996-1.

Correspondence and requests for materials should be addressed to T.G.C.

Chapter 5: Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	2004066	**Title**	Mr
First Name(s)	Anton		
Surname/Family Name	Spadar		
Thesis Title	Understanding the genetic diversity, antimicrobial resistance, and virulence of Klebsiella pneumoniae bacteria		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Scientific Reports		
When was the work published?	Aug 2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I have designed the study, analysed the data, wrote the first draft of the manuscript amd edited manuscript drafts.

SECTION E

Student Signature	
Date	26/09/22

Supervisor Signature	
Date	26/09/22

scientific reports

OPEN

Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains

Anton Spadar[1], João Perdigão[2], Susana Campino[1] & Taane G. Clark[1,3]

The majority of *Klebsiella pneumoniae* (Kp) infections are nosocomial, but a growing number of community-acquired infections are caused by hypervirulent strains (hvKp) characterised by liver invasion and rapid metastasis. Unlike nosocomial Kp infections, hvKp are generally susceptible to antibiotics. Due to the rapid progression of hvKp infections, timely and accurate diagnosis is required for effective treatment. To identify potential drivers of the hypervirulent phenotype, we performed a genome-wide association study (GWAS) analysis on single nucleotide variants and accessory genome loci across 79 publicly available Kp isolates collected from patients' liver and a diverse global Kp dataset (n = 646). The GWAS analysis revealed 29 putative genes ($P < 10^{-10}$) associated with higher risk of liver phenotype, including hypervirulence linked salmochelin *iro* (odds ratio (OR): 29.8) and aerobactin *iuc* (OR: 14.1) loci. A minority of liver isolates (n = 15, 19%) had neither of these siderophores nor any other shared biomarker, suggesting possible unknown drivers of hypervirulence and an intrinsic ability of Kp to invade the liver. Despite identifying potential novel loci linked to a liver invasive Kp phenotype, our work highlights the need for large-scale studies involving more sequence types to identify further hypervirulence biomarkers to assist clinical decision making.

Klebsiella pneumoniae (Kp) is a Gram-negative pathogen increasingly capable of causing severe organ and life-threatening disease. Kp is classified across two main virulence phenotypes, classical (cKp) and hypervirulent (hvKp). CKp is the most common and normally a nosocomial infection, generally occurring among patients with additional co-morbidities[1]. Less common is hvKp, which is characterized by invasive infection within the community setting in otherwise healthy individuals, and with rapid metastatic spread. The typical hvKp presentation involves pyogenic liver abscesses, but also endophthalmitis, meningitis or necrotising fasciitis, all of which are unusual clinical manifestations for cKp. Epidemiologically, hvKp is more common in East and Southeast Asia but is an emerging threat in Europe, particularly when associated with carbapenemase producing clones[1–4].

Biomarkers to differentiate cKp from hvKp are needed to inform diagnostic tests for application by clinical laboratories for optimal patient care and for use in epidemiological surveillance and research studies. However, a complete set of robust biomarkers is not available. Several genetic loci have been identified as virulence factors in Kp, primarily using murine models of infection. These include gene clusters associated with the synthesis of accessory siderophore systems yersiniabactin (*ybt*, *irp1*, *irp2*, and *fyuA*), aerobactin (*iucABCD*, *iutA*), colibactin (*clbA-R*), salmochelin (*iroN*, *iroBCD*), or microcin; mucoidy phenotype regulators (*rmpA* and *rmpA2*), which can up-regulate capsule production; an allantoinase gene cluster; the ferric uptake operon *kfuABC*; and the two-component regulator *kvgAS*, and the K1, K2 and K5 capsular serotypes[1,5–8]. The combination of salmochelin, aerobactin, and *rmpA* is frequently, but not always, linked to the presence of genes from the known Kp virulence plasmids such as pLVPK and pK2044. Some of these may be correlated with hypervirulence[5], but results are inconsistent. In a study of Kp samples from liver abscess samples in East China, only 29% of samples were of

[1]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK. [2]Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Lisboa, Portugal. [3]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. email: taane.clark@lshtm.ac.uk

hypermucoviscous phenotype[9]. Similarly, while the accessory salmochelin locus is frequently found in hvKp samples[6], experimental evidence indicates that aerobactin is the main driver of hypervirulence[10].

Here we analysed the core genome and shared accessory genes of all publicly available Kp samples sourced from the liver (n = 79) and compared them to a large globally diverse public Kp dataset (n = 646) using robust statistical association and cluster analysis methods. Unlike previous studies which leveraged *in vivo* models in either mice (*Mus musculus)* or moths (*Galleria mellonella*) to determine hypervirulence[7,8], we looked at isolates collected from patients' liver, which is a typical clinical presentation site of hvKp. We have found that both acces-sory *iro* and *iuc* loci are strongly associated with liver isolates, and the hypermucoidy associated gene *rmpA* was not linked to hypervirulence. Whilst the analysis revealed new putative loci for the risk of liver phenotype, a minority (19%) of liver isolates did not have any of these markers. Although, the liver phenotype may be subject to misclassification, Kp may have intrinsic ability to colonise the organ, and its genetic underpinning will require a large-scale study to uncover the full repertoire of hypervirulence genes.

Results

Dataset characteristics.

We analysed 79 hvKp isolates defined as samples isolated from patients' liver. These were collected in China (n = 39), Singapore (n = 26), USA (n = 8), Brazil (n = 2) and one sample each from Ecuador, Guadeloupe, South Korea, and Viet Nam (Table 1). Of the 36 sequence types (STs) present in the 79 hvKp samples, ST23 was the most frequent (n = 27) followed by ST86 (n = 9) and ST258 (n = 4). All other STs had two or fewer samples. The 79 hvKp were compared to a large dataset of Kp isolates. This large dataset consisted of two groups: (i) 520 Kp assemblies with similar locations and collection dates to liver isolates, representing the broader genetic landscape of the bacterium; (ii) 126 Kp isolates from three hospitals in Thailand[11], used to assess if our analytical approach was robust, especially to overfitting during data dimensional reduction. Overall, the resulting comparison dataset (n = 646) had samples from 302 different STs among which ST23 (n = 1 7), ST15 (n = 29), ST147 (n = 29), ST11 (n = 25) were the most common.

Association analysis of liver invasive phenotype.

We identified single nucleotide variants (SNVs) in the core genome (5.4 Mbp; 318,458 SNVs, with minor allele frequency (MAF) of 3 Kp isolates). We used a genome-wide association study (GWAS) strategy to identify any SNVs associated with the liver invasive phenotype, adjusting for population structure (Fig. 1A). None of the SNVs associations met our stringent statistical significance level ($P < 10^{-10}$). A similar gene-wide analysis was performed on the presence or absence of accessory loci (n = 15,852), determined from robust assembly of contigs. Whilst the frequency of accessory genes in representative and liver isolates is the broadly correlated (rho = 0.79), the overrepresentation of ST23 (34%) among liver isolates leads to non-linearity (Fig. 2A), which improves when ST23 liver isolates are removed (Fig. 2B) (rho = 0.89). The clustering of isolates based on accessory genome demonstrates that the related genes are linked to ST and not geography, with ST23 being a tight cluster (Fig. S1). We performed the GWAS analysis accounting for this clustering, and found 29 putative genes associated with higher risk of liver phenotype, including known hypervirulence loci *iro* (odds ratio (OR): 29.8) and *iuc* (OR: 14.1), three further metal transport related genes, c-type lysozyme inhibitor (OR: 14.5) and 8 unannotated loci that could not be annotated ($P < 10^{-10}$; Fig. 1B; Table 2). These accessory loci are of lower frequency in representative samples compared to liver isolates, irrespective of inclusion of ST23 (Fig. 2). Of the 79 liver isolates, 15 (19.0%) had none of these 29 putative accessory genes associated with liver invasive phenotype.

Association between identified biomarkers and the rest of the accessory genome.

Having identified 29 accessory genes, including *iro* and *iuc*, with strong potential associations with the hvKp phenotype, we were interested in how they relate to each other i.e., their co-existence. As summarised in a recent review[12], plasmids such as pLVPK, pK2044 and pSGH10 are known carriers of hypervirulence associated genes. Because identified biomarkers do not occur at the same frequency, we hypothesised that they may be on different parts of the hypervirulence plasmids. To test this hypothesis, we performed a cluster analysis of all accessory genes using a *umap* (principal component-like) approach (see "Materials and methods") (Fig. 3). All 29 association loci fell within a cluster of 121 (92 additional) genes (Fig. 3A; Data S3). By focusing on this cluster, *iro* and *iuc* loci are parts of different gene groups (Fig. 3B) consistent with these loci occurring independently of each other, and potentially linked to different hypervirulence plasmids (Fig. S2).

Association between liver invasive phenotype and plasmid replicons.

We evaluated the prevalence of the plasmids identified. Using PlasmidFinder nomenclature, pLVPK, pK2044 and pSGH10 carry IncHI1B(pNDM-MAR) replicons. In pLVPK and pK2044 the replicon sequences are identical. However, based on visual examination of sequences, the first 97nt of pSGH10 are different, while the remaining 472nt are identical to pLVPK and pK2044. In our dataset, 100 isolates had a pLVPK/pK2044 type sequence (20/100; 20.0% liver isolates), while 39 isolates had a pSGH10 type replicon sequence (24/39; 61.5% liver isolates) (Table 3). We observed that pSHG10 type replicons occurred almost exclusively in ST23 isolates (37/39), while a pLVPK/pK2044 type was much more widely distributed, with ST86 (11/100) being most frequent. There was a further variant of IncHI1B(pNDM-MAR) present in single liver isolates from South Korea, which differed from the above variants in the first 120nts. Overall, the most frequent replicon family among liver isolates was IncHI1B(pNDM-MAR) (45/79) followed by IncFIB(K) (16/79).

Liver isolates without identified biomarkers.

Fifteen (19.0%) of the 79 liver Kp isolates did not have the 29 accessory genes associated with the liver phenotype, and included four ST258, two ST1165 and 9 other sequence types. Assuming that the liver invasive phenotype was not misclassified for these 15 samples, we inves-

	Liver samples (n = 79)		Non-liver samples (n = 646)	
Characteristic	**N**	**%**	**N**	**%**
Sequence types				
ST23	27	34	17	3
ST86	8	10	4	1
ST258	4	5	13	2
ST15	–	–	29	4
ST147	–	–	29	4
ST11	1	–	25	4
Other	39	53	529	81
Region				
China	39	49	51	8
Singapore	26	33	1	0
USA	8	10	84	13
South America	4	5	10	2
South Korea	1	1	3	0
Viet Nam	1	1	–	–
Other	0	–	497	77
O types				
O1v1	22	28	147	23
O1v2	39	49	111	17
O2	8	10	152	24
O3	5	6	111	17
Other	5	6	125	19
Carbapenemases				
None	74	94	476	74
KPC-2	4	5	50	8
KPC-3	1	1	25	4
NDM-1	–	–	26	4
Other	–	–	69	11
Aerobactin				
iuc1	45	57	56	9
iuc2	5	6	1	0
iuc3	3	4	11	2
Other	–	–	8	1
None	26	33	570	88
Salmochelin				
iro 1	42	53	43	7
iro 1; iro 3	2	2	1	0
iro 2	5	6	2	0
iro 3	10	13	5	1
Other*	2	2	3	0
None	18	22	592	92

Table 1. Characteristics of study samples. Sequence types (ST); O-types, carbapenemases and siderophore genotypes were determined by Kleborate software; *not reported by Kleborate software.

tigated whether there were any other genes in the accessory genome that differentiated this group from the representative set. By examining differences in allele frequencies between the 15 isolates versus the representative set, we did not find any plausible biomarkers (Figure S3A). We also repeated the core genome GWAS for these 15 samples, but once again there was no SNV which reached the significance cut-off (all $P > 10^{-10}$). It is possible that a combination of accessory genes can predict the phenotype, and we employed nine different machine learning approaches to assess if such a complex gene relationship exists. The imbalance between the 15 hvKp and 646 representative isolates can lead to poor classifier performance in machine learning models, so we ran 100 different datasets with the 15 liver and 15 randomly chosen representative isolates. The resulting predictive accuracy across all approaches was no better than 50% of the random guess (Figure S3B), suggestive that there are no strong predictors of the 19% of liver isolates in our dataset.

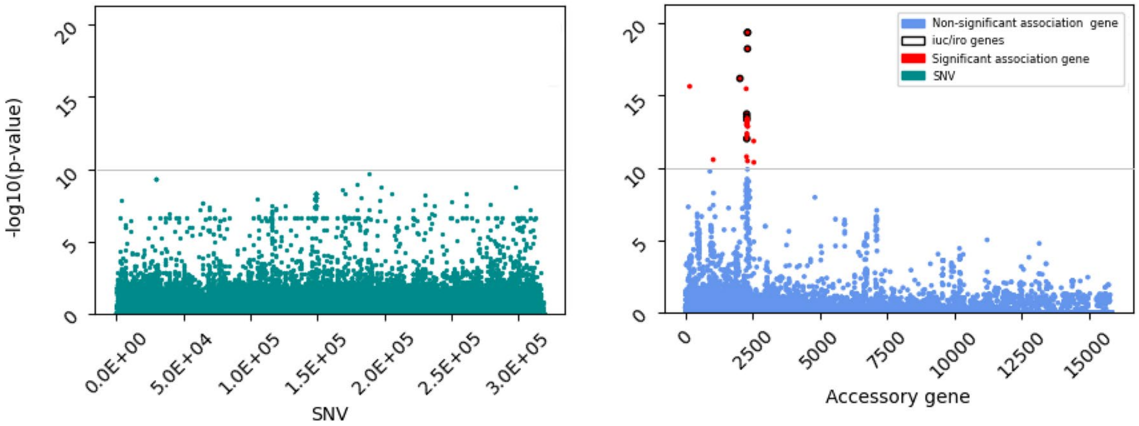


Figure 1. Association analysis of liver versus non-liver against individual genome-wide SNVs (n = 318,458) in the core genome (**A**) and accessory genes (n = 15,852) (**B**), accounting for population structure. Each point represents a result from single SNV or gene, and $P < 10^{-10}$ is the significance threshold.

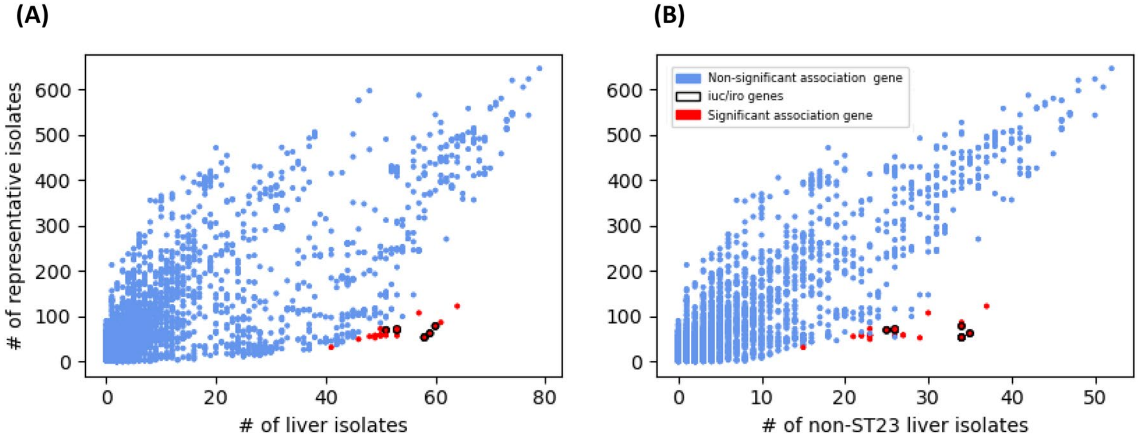


Figure 2. Frequency of accessory genome genes in all liver (**A**) (n = 79) and non-ST23 (**B**) (n = 52) liver isolates versus representative dataset (n = 646). The *iro* and *iuc* outliers are clearly visible. Each point is a gene, and the legend is consistent with Fig. 1.

Discussion

Hypervirulent Kp (hvKp) infections are an emerging global threat with biomarkers needed to differentiate underlying isolates from classical Kp, thereby informing clinical decision making. Previous genetic investigations for hvKp biomarkers have relied on animal models[7,8], where *in vivo* work has identified and focused on both salmochelin *iro* and aerobactin *iuc* loci, sometimes together with genes also present on virulence plasmids. Experimental work has demonstrated that aerobactin is important for Kp survival and growth in human ascites and serum[10]. Additionally, in chicken *E. coli* infection models, both aerobactin and salmochelin have been shown to enhance the colonisation potential of Kp[13,14]. In contrast, our *in-silico* analysis explored 79 Kp samples isolated from the liver, where a liver invasion phenotype is a strong indicator of hvKp. By comparing these isolates with a broader large Kp dataset (n = 646) using a GWAS approach, we found biomarkers on the accessory genome associated with liver hvKp. These markers included *iro* [B] and *iuc* [ABD] loci, as well as *fepA* (a siderophore enterobactin receptor), *IutA* (a ferric aerobactin receptor), *IucA/IucC* (siderophore biosynthesis proteins), and several hypothetical proteins, which serve as candidates for future experiments. *RmpA*, which confers a mucoid phenotype was not found to be associated at our stringent statistical cut-off ($P < 10^{-10}$), but these findings are con-sistent with recent work in carbapenem-resistant Kp[15]. Further, *rmpC* was identified in our GWAS, and ΔrmpC has been shown to maintain the downregulated expression of capsule genes but preserve hypermucoviscosity[16] Another interesting gene is putative c-type lysozyme inhibitor that appears linked to the *iuc* [ABCD] locus. The presence of this gene is potentially associated with the typical clinical manifestation of hvKp in liver and eyes, both organs with high levels of lysozymes[17].

Whilst most of the liver phenotype could be explained through accessory genes, a minority set of isolates did not have any apparent biomarkers. This observation may be explained by phenotypic misclassification where meta data is incorrect, the liver invasive phenotype being intrinsic to Kp, or due to rarely observed genes. Whilst Kp isolate sequence data are likely to be sourced from patients' liver samples, the use of an *in vivo* hypervirulence phenotype can assist phenotypic-genotypic analysis. It is also possible that isolates with known *iuc* and *iro* mark-ers are more likely to be reported compared to samples with undetermined virulence factors. To assess for the

		No. of times gene occurs in isolates			Association	
GeneID	Description	Liver non-ST23 (n = 52)	Liver ST23 (n = 27)	Non-liver (n = 646)	Odds ratio	– log10 P-value
B385452	iroC	34	24	53	29.84	19.33
B385338	iroD	34	24	53	29.84	19.33
B603951	Siderophore enterobactin receptor FepA	35	24	62	23.64	18.19
B362201	iroB	34	26	78	24.68	16.32
B58052	EamA family transporter (*peg-344*)	34	27	86	22.76	15.58
B538146	IS21 family transposase	26	27	57	19.67	15.44
B381713	iucA	26	27	69	14.96	13.48
B385021	rmpC	29	20	52	13.04	13.37
B381836	iucB	26	27	70	14.40	13.28
B597737	Class I SAM-dependent methyltransferase	37	27	122	15.53	13.13
B382206	Ferric aerobactin receptor IutA	26	27	71	14.12	13.11
B382081	iucD	26	27	71	14.12	13.11
B381588	MFS transporter	26	27	71	14.12	13.11
B382762	DM13 domain-containing protein	23	27	57	15.15	13.02
B382654	Hypothetical protein	23	27	57	15.15	13.02
B382870	Hypothetical protein	23	27	57	15.15	13.02
B381162	c-Type lysozyme inhibitor	23	27	58	14.54	12.80
B382331	Hypothetical protein	23	27	58	14.54	12.80
B381271	Peptide deformylase	23	27	58	14.54	12.80
B385565	Hypothetical protein	27	24	58	13.35	12.65
B385675	Hypothetical protein	27	24	58	13.35	12.65
B382547	Hypothetical protein	22	27	57	13.83	12.18
B382440	TetR/AcrR family transcriptional regulator	22	27	57	12.96	12.04
B381960	IucA/IucC family siderophore biosynthesis protein	25	26	69	11.83	11.79
B402327	Tn3 family transposase	21	27	55	12.58	11.67
B380773	Alpha/beta hydrolase	23	27	72	9.90	10.61
B239784	Hypothetical protein	30	27	107	8.24	10.38
B385127	Putative protein	23	23	49	10.78	10.35
B402432	Hypothetical protein	15	26	31	13.66	10.26

Table 2. Relative abundance of accessory genes associated with liver invasive phenotypes identified in Fig. 1B. The DNA sequences for each gene are in Data S2.

presence of sample selection bias, we included a large geographically concentrated dataset from Thai hospitals[11], and consequently found it was not an outlying population in combined analyses with the diverse large global collection. Another limitation is the small number of available hvKp sequences and overrepresentation of the ST23 sequence types. Although, our work is one of the largest hvKp genomic investigations to date, there is a need for larger studies to close knowledge gaps in hvKp epidemiology, pathogenesis, host susceptibility, optimal treatment, and appropriate infection control measures.

Overall, with the increasing prevalence of hvKp strains globally, robust biomarkers of related infection are needed. Our GWAS approach has identified known and novel accessory loci associated with the liver invasive phenotype, some requiring experimental follow-up. It is possible that Kp has an intrinsic ability to invade the liver, requiring larger scale studies to understand the full repertoire of genes underlying hvKp, and thereby improve clinical decision making.

Materials and methods

Dataset.

We identified potential hvKp samples with sequencing data by searching the NCBI Isolates Browser[18] (November 2021) using key words "liver" and "hepa". Metadata of positive hits were manually examined to confirm a likely liver invasive phenotype. We did not identify any samples isolated from endophthalmitis, which is an infrequent manifestation of hvKp. The search resulted in 79 samples, of which 31 had sequencing

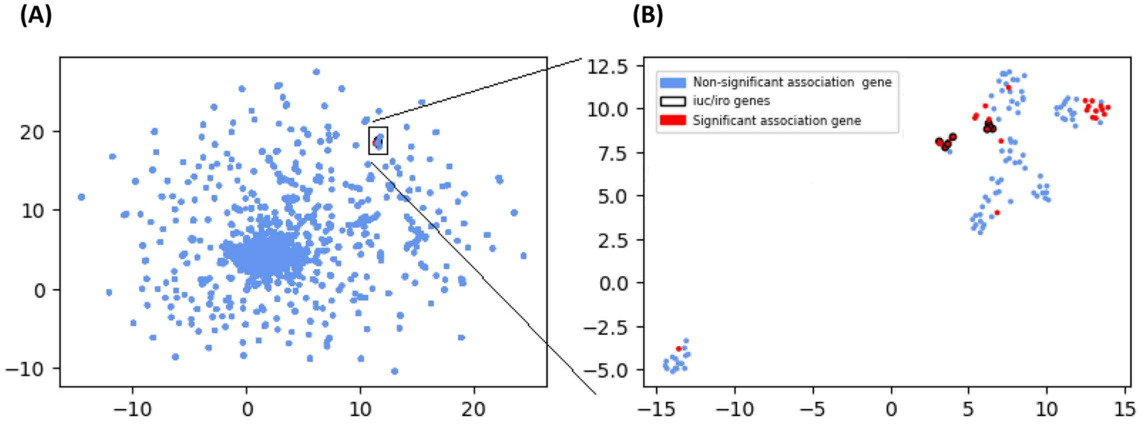


Figure 3. Cluster analysis of accessory genes. (**A**) Projection of genes presence/absence matrix into a *umap* 2-dmiensional view; (**B**) Structure of the *iro* and *iuc* containing gene cluster in (**A**). The liver phenotype genes (Table 2) are visible both in (**A**) and in greater detail in (**B**) for which the dimensional reduction algorithm was re-ran with subset of genes in (**A**). The axes are dimensionless. Each point is an accessory gene.

Replicons	Total	STs (no. isolates)	Countries (no. isolates)	From liver	With iuc	With iro
IncHI1B(pNDM-MAR) [pLVPK/pK2044 type]	100	ST86 (11), ST23 (6), ST15 (6), ST14 (5)	China (26), Thailand (24), Singapore (8), USA (5), United Kingdom (4)	20 (20.0%)	48 (48.0%)	39 (39.0%)
IncHI1B(pNDM-MAR) [pSHG10 type]	39	ST23 (37), ST1941 (1), ST152 (1)	China (15), Singapore (12), Thailand (8)	24 (61.5%)	39 (100%)	36 (92.3%)

Table 3. Prevalence of IncHI1B(pNDM-MAR) plasmid replicons.

reads and 48 were sequence assemblies. We assembled the sequencing reads for all samples using Unicycler v0.4.8[19] with a quality check performed using Busco software (v4)[20] to ensure > 95% completeness and < 5% fragmentation of genes in the gammaproteobacteria_odb10 gene set. For consistency of downstream analysis, all samples were re-annotated with prokka software (v1.14.6)[21] using the Klebsiella genus database[22] and default settings.

The 79 hvKp samples were complemented by 520 randomly selected assemblies also from the NCBI Isolates Browser. However, before the random selection we identified groups of isolates matching by location, isolation source and create date. We removed all but one representative isolate from each group, to minimize bias from large, localized studies. These randomly chosen samples may have characteristics of hvKp, but they provide an important comparison for establishing if a set of genes is more common in hvKp compared to those in the broader population. We also enriched our dataset with a further 126 samples[11] isolated from three hospitals in Thailand, to evaluate the impact of samples chosen from a small geographic area with a diversity of STs and assess the robustness of analysis. If our methods are prone to generating bias, we would expect this dataset to stand out, but it did not (see Fig. S1). The comparison dataset of 646 isolates consisted of 302 different STs with ST15 (n = 29), ST147 (n = 29), ST11 (n = 25) being the most common. Kleborate software (v2.1.0)[23] was used to profile the isolates' virulence and ST (Data S1).

Analysis. The genes from all assemblies were clustered in a reference independent manner. The Kp core genome was identified as those genes which are not accessory. To identify a core genome, BLASTn (v2.9.0)[24] with word-size 20 was used to find and remove genes that shared > 90% identity, were within 20% of median length of all such genes, and were present in > 90% of samples. A sensitivity analysis performed with alternative parameters produced similar results. This approach identified a conserved core gene set which was removed. For the remaining genes we performed an all versus all BLASTn search with word-size 11. We assigned genes to groups based on > 60% identity between any two genes intra group and < 20% length difference from median gene length intra group. The input for subsequent analysis was a 15,852 × 725 matrix with rows as gene groups and columns as samples, where individual cells are a binary value with one indicating that sample contains a gene from the group, zero otherwise. Genes were aligned using MAFFT software (v7.467)[25] and the resulting alignment files transformed into a 318,458 × 725 python matrix, where rows are individual SNVs and columns are isolates.

Logistic regression models were used to find associations between the liver phenotype and SNVs or presence of accessory genes. These models included principal components for the population structure, and were implemented using statsmodels software (v0.13.0)[26]. The projection of the dataset into two dimensions was performed using the umap library (v0.5.1)[27] in python using "hamming" distance. Clusters were determined using DBSCAN[28] as implemented in sklearn (v0.24.2)[29]. Machine learning analysis was performed using sklearn functions to identify predictors of the liver phenotype. Plasmid replicons were identified using PlasmidFinder software (v2.1.1) with default settings[30]. The scripts for accessory genome construction are available at https://

github.com/AntonS-bio/accessoryGenomeBuilder. The analysis scripts are available at https://github.com/AntonS-bio/KpHypervirulence.

Ethics approval and consent. No ethics approvals were required as all data is publicly available.

Data availability

All data used in this work is publicly available in NCBI database (https://www.ncbi.nlm.nih.gov/). A list of isolates is in Data S1. Analysis scripts are available at https://github.com/AntonS-bio.

Received: 15 March 2022; Accepted: 3 August 2022
Published online: 11 August 2022

References

1. Russo, T. A. & Marr, C. M. Hypervirulent *Klebsiella pneumoniae*. *Clin. Microbiol. Rev.* **32**, 31092506 (2019).
2. Zhang, Y. *et al.* High prevalence of hypervirulent *Klebsiella pneumoniae* infection in China: Geographic distribution, clinical characteristics, and antimicrobial resistance. *Antimicrob. Agents Chemother.* **60**(10), 6115–6120 (2016).
3. Struve, C. *et al.* Mapping the evolution of hypervirulent *Klebsiella pneumoniae*. *MBio* **6**, 4 (2015).
4. European Centre for Disease Prevention and Control. *Risk Assessment: Emergence of Hypervirulent Klebsiella pneumoniae ST23 Carrying Carbapenemase Genes in EU/EEA Countries*. (2021).
5. Catalán-Nájera, J. C., Garza-Ramos, U. & Barrios-Camacho, H. Hypervirulence and hypermucoviscosity: Two different but com-plementary *Klebsiella* spp. phenotypes?. *Virulence* **8**, 1111–1123 (2017).
6. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumo-niae*, an urgent threat to public health. *Proc. Natl. Acad. Sci. USA* **112**(27), E3574–E3581 (2015).
7. Li, G. *et al.* Identification of hypervirulent *Klebsiella pneumoniae* isolates using the string test in combination with *Galleria mel-lonella* infectivity. *Eur. J. Clin. Microbiol. Infect. Dis.* **39**(9), 1673–1679 (2020).
8. Russo, T. A. *et al.* Identification of biomarkers for differentiation of hypervirulent *Klebsiella pneumoniae* from classical *K. pneu-moniae*. *J. Clin. Microbiol.* **56**, 9 (2018).
9. Qu, T. *et al.* Clinical and microbiological characteristics of *Klebsiella pneumoniae* liver abscess in East China. *BMC Infect. Dis.* **15**, 1 (2015).
10. Russo, T. A., Olson, R., MacDonald, U., Beanan, J. & Davidsona, B. A. Aerobactin, but not yersiniabactin, salmochelin, or entero-bactin, enables the growth/survival of hypervirulent (hypermucoviscous) *Klebsiella pneumoniae ex vivo* and *in vivo*. *Infect. Immun.* **83**(8), 3325–3333 (2015).
11. Loraine, J. *et al.* Complement susceptibility in relation to genome sequence of recent *Klebsiella pneumoniae* isolates from Thai hospitals. *MSphere.* **3**, 6 (2018).
12. Yang, X., Dong, N., Chan, E. W. C., Zhang, R. & Chen, S. Carbapenem resistance-encoding and virulence-encoding conjugative plasmids in *Klebsiella pneumoniae*. *Trends Microbiol.* **29**(1), 65–83 (2021).
13. Gao, Q. *et al.* The avian pathogenic *Escherichia coli* O2 strain E058 carrying the defined aerobactin-defective iucD or iucDiutA mutation is less virulent in the chicken. *Infect. Genet. Evol.* **30**, 267–277 (2015).
14. Gao, Q. *et al.* Roles of iron acquisition systems in virulence of extraintestinal pathogenic *Escherichia coli*: Salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. *BMC Microbiol.* **12**, 143 (2012).
15. Shankar, C. *et al.* Aerobactin seems to be a promising marker compared with unstable RmpA2 for the identification of hypervirulent carbapenem-resistant *Klebsiella pneumoniae*: *In silico* and *in vitro* evidence. *Front. Cell. Infect. Microbiol.* **11**, 1 (2021).
16. Walker, K. A., Treat, L. P., Sepúlveda, V. E. & Miller, V. L. The small protein rmpd drives hypermucoviscosity in *Klebsiella pneu-moniae*. *MBio* **11**(5), 1–14 (2020).
17. Ragland, S. A. & Criss, A. K. From bacterial killing to immune modulation: Recent insights into the functions of lysozyme. *PLoS Pathog.* **13**, e1006512. https://doi.org/10.1371/journal.ppat.1006512.g001 (2017).
18. National Library of Medicine. *The NCBI Pathogen Detection Project*. (2016).
19. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**(6), e1005595 (2017).
20. Seppey, M., Manni, M. & Zdobnov, E. M. *BUSCO: Assessing Genome Assembly and Annotation Completeness* 227–245 (Humana Press Inc., 2019).
21. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**(14), 2068–2069 (2014).
22. Ehrlich, R. *Prokka Database Maker*. (2019). https://github.com/rehrlich/prokka_database_maker. Accessed 24 May 2014.
23. Lam, M. M. C., Wick, R. R., Wyres, K. L. & Holt, K. E. Genomic surveillance framework and global population structure for *Klebsiella pneumoniae*. *Biorxiv.* https://doi.org/10.1101/2020.12.14.422303 (2020).
24. Agarwala, R. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **44**(D1), D7-19 (2016).
25. Katoh, K. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**(14), 3059–3066 (2002).
26. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *9th Python in Science Conference* (2010). http://statsmodels.sourceforge.net/.
27. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. (2018). http://arxiv.org/abs/1802.03426.
28. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited. *ACM Trans. Database Syst.* **42**(3), 1–21 (2017).
29. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**(85), 2825–2830 (2011).
30. Carattoli, A. *et al. In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**(7), 3895–3903 (2014).

Acknowledgements

SC was funded by the Medical Research Council UK (Grant No. MR/M01360X/1) and BBSRC UK (BB/R013063/1). TGC was supported by the Medical Research Council UK (Grant No. MR/K000551/1, MR/M01360X/1, MR/N010469/1, MR/R020973/1) and BBSRC (BB/R013063/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

A.S. and T.G.C. designed the study, and A.S. analysed the data under the supervision of J.P., S.C. and T.G.C. A.S. wrote the first draft of the manuscript, with contributions from J.P., S.C. and T.G.C. All authors have edited manuscript drafts and agreed on the contents of the final version. All authors have consented to the publication of this manuscript.

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/s41598-022-17995-2.

Correspondence and requests for materials should be addressed to T.G.C.

Reprints and permissions information is available at www.nature.com/reprints.

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

Chapter 6: Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	2004066	**Title**	Mr
First Name(s)	Anton		
Surname/Family Name	Spadar		
Thesis Title	Understanding the genetic diversity, antimicrobial resistance, and virulence of Klebsiella pneumoniae bacteria		
Primary Supervisor	Prof. Taane Clark		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Genome Medicine
Please list the paper's authors in the intended authorship order:	Anton Spadar, João Perdigão, Susana Campino, Taane G. Clark
Stage of publication	**Revised**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I have designed the study, analysed the data, wrote the first draft of the manuscript and edited the manuscript drafts.

SECTION E

Student Signature	
Date	26/09/22

Supervisor Signature	
Date	26/09/22

Title: Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains.

Anton Spadar[1], João Perdigão[2], Susana Campino[1], Taane G. Clark [1,3]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

[2] Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Lisboa, Portugal

[3] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

Corresponding author:

Prof. Taane G. Clark

Department of Infection Biology,

London School of Hygiene and Tropical Medicine

Keppel Street, London WC1E 7HT

United Kingdom

taane.clark@lshtm.ac.uk

ABSTRACT

Klebsiella pneumoniae (Kp) Gram-negative bacteria cause nosocomial infections and rapidly acquire antimicrobial resistance which makes it a global threat to human health. It also has a comparatively rare hypervirulent phenotype that can lead to severe disease in otherwise healthy individuals. Unlike classic Kp, canonical hypervirulent strains usually have limited antimicrobial resistance. However, after initial case reports in 2015, carbapenem resistant hypervirulent Kp has increased in prevalence. It is now endemic in China, but there is limited understanding of its prevalence in other geographical regions. Here, we examined the largest collection of publicly available sequenced Kp isolates (n=13,178), containing 1,603 different sequence types (e.g., ST11 15.0%, ST258 9.5%), and 2,174 (16.5%) hypervirulent strains. We identified and analysed 3,034 unique plasmid replicons to understand the epidemiology and transmission dynamics of carbapenem resistant hypervirulent Kp (n=1,028, 7.8%). We identified several outbreaks globally including one involving ST11 strains in China and another of ST231 in Asia centred on India, Thailand, and Pakistan. There was evidence of global flow of Kp, including across multiple continents. In most cases, clusters of isolates are the result of hypervirulence genes entering classic Kp strains instead of carbapenem resistance genes entering canonical hypervirulent strains. Overall, our analysis demonstrates the utility of genomic sequencing of Kp to monitor global epidemiology carbapenem resistant and hypervirulent strains. With the growing adoption of monitoring technologies, including in geographical regions with gaps in data and knowledge (e.g., Sub-Saharan Africa), the identification of the spread of antimicrobial resistance will inform infection control globally.

KEYWORDS: AMR, carbapenem, hypervirulence, klebsiella

Word count: 250

BACKGROUND

Klebsiella pneumoniae (Kp) is a Gram-negative pathogen increasingly capable of causing severe organ and life-threatening disease. Kp is classified into two main virulence phenotypes: hypervirulent (hvKp) and classical (i.e., non-hvKp). The classic Kp is the most common and normally a nosocomial infection that occurs in patients with additional co-morbidities (1). Less common is hvKp which is characterized by invasive infection within the community setting in healthy individuals and a rapid metastatic spread. Epidemiologically, hvKp is more common in East and Southeast Asia, but it is also an emerging threat in Europe. HvKp associated with carbapenemase producing clones is particularly concerning (1–4) as it will make infection control more difficult. It is therefore important to use genomic data and analyses from local, regional, and global studies to monitor its emergence and spread.

The main driver of carbapenem resistance in Kp is the acquisition of genes that encode carbapenemases such as KPC, NDM, some variants of OXA and others (5,6). The encoding genes are usually located on small mobile genetic elements such as transposons or insertion sequences. These elements are themselves usually embedded in plasmids which mediate the transfer of genes between bacterial cells via horizontal gene transfer (7–9).

Similarly, hypervirulence is normally associated with aerobactin (*iuc*) and salmochelin (*iro*) gene loci carried on plasmids. The *iro* and *iuc* loci are frequently accompanied by additional genes (e.g., *rmpA*, *rmpA2*, and *rmpC*) associated with a hypermucoviscous capsule phenotype (10,11). K1 and K2 are the dominant capsular genotypes among hvKp (12). The *iuc* and *iro* loci usually reside on large (>200kbp) plasmids but are occasionally located on chromosomes via integrated chromosomal elements (1).

Since reports of carbapenem resistant hypervirulent Kp (CRhvKp) in a tertiary hospital in Beijing in 2015 (13), the cases have steadily increased in number and geographic range (14–17). The incidence of CRhvKp in China was estimated at ~13% based on the presence of *rmpA*, *rmpA2* or *iutA* biomarkers (15), but estimates vary regionally within the country. Another large-scale study in China reported that 36% of screened carbapenem resistant (CRKp) strains carried the hypervirulence associated siderophore aerobactin locus (17). Other regional studies focused on Singapore (18), India (19) and USA (20), but understanding of global landscape is limited. With increased human mobility and volumes of trade and tourism, knowledge of the global genomic diversity of plasmids underlying CRhvKp strains will assist with monitoring the emergence and spread of CRhvKp infections and inform their clinical management.

In this study, we analysed the plasmid replicons as well as carbapenemase and siderophore encoding genes across 12,468 geographically diverse Kp isolates to understand the movement of hypervirulence and antimicrobial resistance genes on plasmids and their convergence in CRhvKp. Despite the limitations of the convenience sampling, we found growing prevalence of CRhvKp assigned to nine clusters, seven of which are statistically robust. These clusters included two previously reported outbreaks identified in China (15–17,21); however, they also include additional outbreaks with one spanning Asia, Africa, and Europe. Our findings demonstrate the utility of large-scale data for understanding the epidemiology of CRhvKp, their emergence and spread; we present a first updateable categorisation of CRhvKp isolates, with potential utility for clinical and surveillance investigations.

METHODS

All publicly available assemblies labelled as Kp (n=13,178) in the NCBI RefSeq database (as at September 2021) were downloaded (22). Kleborate software (v2.2.0) was used to confirm species, and identify sequence types (ST), find antimicrobial resistance (AMR) and virulence genes, and determine capsular and O-antigen types (23) **(Fig S1, Data S1)**. Isolates were classified as hypervirulent Kp (hvKp) genotype if they contained either aerobactin (*iuc*), salmochelin (*iro*) or both gene loci (10,11). Carbapenem resistance can occur without carbapenemase encoding genes, but we conservatively defined CRKp genotype as isolates that carry a carbapenemase encoding gene. Plasmid replicons were identified using PlasmidFinder software (v2.1.1) (24) with default cut-offs (>=60% coverage and >=90% identity). The isolate data included complete, contig and scaffold levels assemblies.

PlasmidFinder compares assemblies to a database of nucleotide sequences of genes encoding replication control and initiation proteins. The assembly sub-sequences (not necessarily whole contigs) which pass similarity thresholds are returned by the software. The returned sub-sequences were summarised in a binary data matrix (1 = present, 0 absent), where the rows represented individual isolates, and columns (n=3,034) represented unique replicons nucleotide sequence. This binary matrix was used to calculate the r*ussellrao* distance between pairs of isolates as implemented in sklearn software (v0.24.2) (25).

We used the UMAP (v0.5.1) algorithm to project the full 3,034-dimensional binary matrix into 2-dimensional space (25–27). UMAP aims to project multidimensional data into fewer dimensions while preserving some global and local data topology. The approach has similarity to principal components analysis (PCA) and multi-dimensional scaling (MDS or PCoA). The former uses a covariance matrix, and the latter uses a matrix of pairwise distances. Neither PCA nor MDS are intended for binary matrices, though MDS can accommodate ordinal data (28,29). UMAP belongs to family of manifold learning algorithms which are usually better at dimensional reduction of binary data (30–32). We applied HDBSCAN software (v0.81) (min_cluster_size=10 and cluster_selection_epsilon=0.5) to the 2D projection generated by UMAP to identify clusters of isolates (33). HDSCAN is a density-based clustering algorithm which does not require all data points to be a part of a cluster, and we have labelled such isolates as “unassigned”.

Unlike PCA and MDS, UMAP is a stochastic algorithm; so we assessed the robustness of clustering based on UMAP projection. For this purpose, we created a Monte Carlo simulation by repeating the UMAP projection and cluster detection 500 times. The clusters identified in each of 500 iterations formed columns of a data matrix which was itself embedded using UMAP with a *hamming* distance metric. The clusters in this projection were again determined using HBSCAN **(Fig S1)**. These latter clusters aggregate 500 individual runs and to determine the consistency of aggregated versus each underlying cluster, we calculated chi-squared tests, implemented in sklearn software. We labelled the aggregate cluster as robust if less than 1% of -log$_{10}$(p-value) was below 20 **(Fig S1)**

We also compared replicons found in our dataset to those characterised in 35 historic bacterial isolates (34) (“Murray collection”) sourced between 1920 and 1949 and classified as Kp based on Kleborate software typing.

RESULTS

Isolates and genotypes

While all isolates (n=13,178) were Kp according to metadata, Kleborate screening identified them as Kp (n=11,820; 90%), *K. quasipneumoniae* (n=604, 5%), *K. variicola* (n=428; 3%), *K. aerogenes* (n=299; 2%) and other subspecies of *Klebsiella* (n=27; 0.2%) i.e., some isolates may have incorrect species data in the NCBI database **(Fig S1)**. We did not restrict the analysis to *K. pneumoniae sensu stricto*, but we removed 710 isolates in which PlasmidFinder did not identify any replicons. These removed isolates covered 331 STs, with ST3910 (n=21) being most frequent. They also had a high abundance of non Kp *sensu stricto* (*K. quasipneumoniae* n=140, 19.7%; *K. aerogenes* n=115, 16.2%; *K. variicola* (n=109, 15.4%). Notably, these isolates had limited AMR carriage with only 9 isolates carrying carbapenemase encoding genes, among which *bla*$_{OXA-48}$ was the most frequent (n=4). There were also few fluroquinolone *gyrA* mutations (n=65/710) and aminoglycoside resistance enzymes (n=35/710). Only four isolates carried an aerobactin locus; however, 20 had a truncated *iro3* locus and 109 *K. aerogenes* samples had a chromosomally carried salmochelin locus.

The final isolate dataset (n=12,468) contained 1,603 different STs (based on 7 chromosomal loci) with the most frequent being ST11 (15%), ST258 (10%), ST15 (4%), ST512 (4%) ST307 (3%) and ST147 (3%) **(Fig S1)**. The number of hypervirulent strains was 1,881 (15%). The two canonical hypervirulent strains ST23 and ST86 accounted for 1.8% and 0.7%, respectively. The study had global representation, with isolates from 103 countries, including China (20%), USA (14%), Italy (7%), United Kingdom (6%), Thailand (4%), and Germany (4%); however only 2% of all isolates were from Sub-Saharan Africa. While the convenience nature of the sampling may make the dataset unsuitable for the estimation of the prevalence of CRhvKp genotypes, the large sample size presents a likely accurate assessment of regional and temporal trends.

Replicons

An important cornerstone of our analysis is the distinction between replicon family or name (e.g., IncFII(K), IncFIB(K) and ColRNAI) and the underlying unique nucleotide sequences of the replicons belonging to specific family. For example, across the 12,468 isolates there were 3,034 unique replicon nucleotide sequences, of which 1,096 occurred in more than one isolate. The most frequent replicon nucleotide sequence was the PlasmidFinder reference version of IncFIB(K), which occurred in 3,123 isolates (24%). However, the replicons from an IncFIB(K) family occurred 7,052 times with eight distinct sequences occurring in over 100 Kp isolates **(Fig 1)**. The family and nucleotide sequence diversity are summarised in **Data S2**.

The average number of replicons per isolate was 4.6 (range: 1 to 29). A total of 6,783 isolates (55%) with identified replicons had carbapenemase encoding genes, and 1,881 (15%) had either *iuc* or *iro* loci. A total of 1,028 (8%) isolates with identified replicons had both carbapenemase encoding and hypervirulence (CRhvKp) loci **(Fig S1)**. Therefore, the numbers of isolates determined to be CRhvKp, CRKp (non- hvKp) and hvKp (non-CRKp) were 1,028 (8%), 5,755 (46%), and 853 (7%), respectively **(Fig S1)**. The majority of the CRhvKp were from China (n=696) representing 27% of isolates from that country with isolates collected in Russia (n=64), Thailand (n=44), Italy (n=39) and India (n=25) being next most common.

To examine the geographical and temporal trends of carbapenem resistant hypervirulent genotypes, we assessed the replicon-based clustering of CRKp, hvKP and CRhvKp isolates across 695 plasmid replicon nucleotide sequences that were present in multiple CRKp, hvKp and CRhvKp isolates **(Fig S1)**. UMAP based analysis of population structure revealed that clustering was not driven solely by geography or ST **(Fig 2)**. This contrasts with chromosomal-based cluster analysis, which revealed clustering of isolates by ST (data not shown). We used HDBSCAN clustering algorithm on the two-dimensional embedding of CRhvKp isolates (n=1,028) plasmids replicons and identified 9 groups (Clusters A – I; **Table 1**, **Fig 2**), with 79 (8%) isolates not assigned to any group (see **Data S3** for all assignments). Of these 9 clusters, only two (B and E) did not have strong statistical support of their robustness.

Replicon clusters

Cluster A (n=560, **Table 1**) consisted mainly isolates from China and accounted for the majority of CRhvKp isolates from that country (n=517/696). This cluster also included isolates with travel links to China (35). The first isolate from Cluster A was collected in 2012 with frequency increasing over time **(Fig 3)**. ST11 was the dominant sequence type in this cluster (n=499/560), and this cluster also contained majority of ST11 CRhvKp isolates (n=499/582). Hypervirulence was driven by *iuc1* in nearly all isolates (n=554/560) and 168 isolates additionally had *iro1* (77 truncated or incomplete) following the nomenclature established previously (36). The carbapenemase encoding genes were dominated by *bla*$_{KPC-2}$ (n=516/560), with *bla*$_{OXA-48}$ the next most common (n=16) **(Table 1, Data S3)**.

The dominant replicons in Cluster A were ColRNAI (n=943), which occurred multiple times in most isolates, followed by repB (n=555), IncFII(pHN7A8) (n=501), IncHI1B(pNDM-MAR) (n=489) and IncR (n=486) **(Data S2)**. By examining assembly contigs for co-presence of carbapenemase, virulence genes and replicons, we were able to link replicon and carbapenemase genes in 74 isolates and replicon and siderophore genes in 341 isolates **(Table 1)**. No carbapenemase or hypervirulence linked siderophore genes were found on chromosomes. In cases where siderophores were located on the same contig as replicons, the vast majority (n=290/343) had a IncHI1B(pNDM-MAR) replicon **(Data S4)**. Of these, 125 contigs had both repB and IncHI1B(pNDM-MAR) replicons. In addition, 42 contigs had simultaneously repB replicon and siderophores. The majority (n=52/76) of contigs which carried replicon and carbapenemase genes had an IncFII(pHN7A8) replicon, and 49 of these also had an IncR replicon. Two contigs had both siderophores and carbapenemase genes. The first one had repB, IncFIB(pKPHS1) and IncHI1B(pNDM-MAR) replicons. The second had repB and IncHI1B(pNDM-MAR). Both contigs carried *bla*$_{KPC-2}$ and *iuc1* loci.

Cluster B (n=83) consisted mainly of isolates from China (n=81/83) and ST11 sequence type (n=74/83). This cluster did not have strong statistical support for its robustness, and based on visual examination of results **(Fig 2)** it is related to, but distinct from, Cluster A isolates. In particular, the isolates in Cluster B lack repB and IncHI1B(pNDM-MAR) replicons, characteristic of Cluster A.

Cluster C (n=82) is interesting due to its geographic diversity with 32 isolates sourced from Russia, 16 from China, 9 from Egypt, and 8 from Germany. There were 15 STs of which ST147 (n=32; 39%) and ST395 (n=24; 29%) were most frequent. While ST147 had broad geographic distribution the majority of ST395 isolates in this cluster were collected in Russia (n=20/24; 83%). The most frequent carbapenemases were *bla*$_{OXA-48}$ (n=51) and *bla*$_{NDM-1}$ (n=23), again without strong geographic links. Five isolates had both genes. More broadly, 78 isolates had either a *bla*$_{OXA}$ or *bla*$_{NDM}$ gene, with *bla*$_{KPC-2}$ (n=2) and *bla*$_{VIM-1}$ (n=1) accounting for the rest **(Data S2)**. The dominant hypervirulence siderophore was *iuc1* (n=78/82), while an unassigned salmochelin lineage was present in two isolates. Despite geographic and ST diversity, three replicons accounted for majority of isolates: Col(pHAD28) (n=91), IncHI1B(pNDM-MAR) (n=80), and IncFIB(pNDM-Mar) (n=79) **(Table 1)**. Out of 14 isolates in which a replicon and *iuc1* locus were on the same contig, ten had both IncFIB(pNDM-Mar) and IncHI1B(pNDM-MAR) replicons, and a further two had IncHI1B(pNDM-MAR). Replicons linked to carbapenemase encoding genes (n=16) were diverse, and IncL linked to *blaOXA-48* (n=4) was the most frequent association.

Cluster D (n=61) consisted of isolates mainly from Southeast Asia (Thailand, n=27; India, n=17; Pakistan, n=8). The first isolate was collected in Malaysia in 2013 while the most recent six isolates were collected in India in 2019. Nearly all isolates belong to ST231 (n=60/61) which shares only two MLST alleles with canonical hypervirulent ST23. The dominant carbapenemase was *bla*$_{OXA-232}$ (n=59/61) with *bla*$_{OXA-181}$ and *bla*$_{OXA-48}$ in other two isolates. Unusually, hypervirulence was driven by *iuc5* (n=60/61), and accounts for nearly all *iuc5* carrying CRhvKp (n=60/75). Salmochelin (*iro1*) was only present in a single isolate. This cluster had near universal carriage of seven replicons: Col440I, IncFIB(pQil), IncFII(K), IncFIA, IncFII(pAMA1167-NDM-5), ColKP3, and Col(pHAD28) **(Table 1)**. IncFII(pAMA1167-NDM-5) was unique to this cluster and IncFIA only occurred in nine further CRhvKp isolates. We were able to link replicons to a carbapenemase gene in 60 isolates, as well as to a siderophore in 8 isolates. Neither siderophore nor carbapenemase genes occurred on chromosomes. In all cases, *bla*$_{OXA-232}$ was linked to the ColKP3 replicon, while *iuc5* was co-located on a contig with IncFIA and IncFII(pAMA1167-NDM-5).

Cluster E (n=50) contained isolates mainly from Asia (China, n=16; Singapore, n=14; Thailand, n=13), but unlike Clusters A and B, this one also had five samples from three European countries (Latvia, n=2; Greece, n=2; France, n=1). The first isolates were collected in 2013 in Singapore (n=7) and the most recent isolated in China in 2020. Sixteen isolates were of the canonical hypervirulent strain ST86 and collected in six countries. This cluster also had ST65 (n=14) and ST23 (n=4) isolates. The former shares only two alleles with ST86. The dominant siderophores in this cluster were *iuc1* (n=44) and *iro1* (n=40) with 40 isolates carrying both. While *bla*$_{KPC-2}$ was the most frequent carbapenemase gene (n=32), some isolates (n=13) carried *bla*$_{OXA-232}$ just like the isolates from Cluster B. All these *bla*$_{OXA-232}$ carrying isolates were collected as part of the same Thai study in 2016; however, they belonged to 8 different STs (37). Unlike Clusters A and B, this cluster had only two near universal replicons: repB (n=50) and IncHI1B(pNDM-MAR) (n=49). We were able to link replicon to carbapenemase genes in 19 and to a siderophore in 32 isolates. Nearly all linked *iuc* loci were located on a contig with either IncHI1B(pNDM-MAR) (n=4), repB (n=1) or both (n=22) (**Data S4**). The *bla*$_{OXA-232}$ gene was co-located with ColKp3 replicon in all *bla*$_{OXA-232}$ carrying isolates, like Cluster B. The replicons linked to *bla*$_{KPC-2}$ were much more diverse. Out of 6 contigs three had a IncFII(K) replicon, two IncFII(pHN7A8), and one IncX6. None of the examined contigs carried both carbapenemase and hypervirulence genes, nor were any of these genes on chromosomes.

Cluster G consists of 32 isolates collected every year between 2015 and 2020 with majority isolated in China (n=26) **(Table 1)**. While this cluster includes two ST23 and one ST11 isolates, it is dominated by ST15 (n=26/32). The main siderophore locus was *iuc1* (n=32/32) with two isolates concurrently carrying *iro1*. The dominant carbapenemase was *bla*$_{OXA-232}$ (n=27/32) with *bla*$_{NDM-1}$ (n=3/32) next most common. The most interesting aspect of this cluster was the diversity of frequent replicons: repB(n=31), ColRNAI(n=31), IncHI1B(pNDM-MAR)(n=30), IncFIB(pKPHS1)(n=28), IncFII(K)(n=28), ColKP3(n=28), and Col(pHAD28)(n=28). We linked replicon to carbapenemase in 32 and to a siderophore in 7 isolates. The *iuc1* hypervirulent locus was linked to IncHI1B(pNDM-MAR) (n=6) and blaOXA-232 was linked to ColKP3(n=28) as in cluster D, but that cluster has *iuc5* linked to IncFIA and IncFII(pAMA1167-NDM-5). Part of this cluster has been described previously (21), reinforcing the robustness of our approach.

Salmochelin carrying K. aerogenes

Surprisingly, nearly all *K. aerogenes* isolates (n=288/299) carried a salmochelin locus, but none had aerobactin. In all 41 complete *K. aerogenes* assemblies this locus was located on chromosome (contig > 4,500,000nt). These salmochelin genes had nucleotide identity between 74% and 86% to *Kp sensu stricto* salmochelin genes. The nearest amino acid sequences outside *K. aerogenes* were *iroB, iroC, iroB* and *iroN* in *Enterobacter oligothropicus* with 93%, 90%, 81% identity, respectively. A small portion of *K. aerogenes* isolates had both salmochelin and carbapenemase genes (n=40), among which carbapenemase genes $bla_{KPC\text{-}2}$ (n=13) and $bla_{OXA\text{-}48}$ (n=6) were the most common. Most of the *K. aerogenes* without carbapenemase genes (n=231/248) did not have any extended-spectrum beta-lactamase (ESBL) encoding loci, despite most of them (n=171/231) being collected after 2010 – a period in which ESBL encoding genes are common in Kp. PlasmidFinder identified replicons in 174 salmochelin carrying *K. aerogenes*. Their replicons formed a cluster of samples consisting mainly of USA isolates (n=73) with a few from Germany (n=9), Lebanon (n=6) and other countries. Apart from one small study (38) we believe this is the first major report of the widespread presence of salmochelin in *K. aerogenes*.

Comparison to replicons of historic isolates

Plasmid sequences of 35 Kp isolates from the historic Murray collection were identified and compared to the 3,034 unique replicons in our whole collection (n=12,468). Remarkably, there was a substantial overlap between these unique replicon sequences. For example, the same *repB* sequences occurred in 1,124 of all isolates and in 19 Murray Collection isolates. Three further replicon sequences [IncFIB(K), IncFII(pKP91) and IncHI1B(pNDM-MAR)] that occurred once each in the Murray Collection occurred in over 240 general isolates **(Data S5)**. The co-existence of multiple variants of identical replicons up to 90 years ago requires further investigation into the evolution of plasmid replicon sequences including their mutation rates and any selective pressure that are currently unknown.

IncHI1B(pNDM-MAR) replicon

The Cluster A version of *repB* gene differs from PlasmidFinder's IncHI1B(pNDM-MAR) by 4nt and from the closest sequence in our collection by 3nt out of 570nt. The latter sequence occurs frequently (n=395) in our whole dataset **(Data S2)**. More importantly, the first 96nt of this replicon's sequence (those preceding *repB* sequence) are almost unique to Cluster A. These leading 96nt have no similarity using BLAST (word size 7) (39) to any other IncHI1B(pNDM-MAR) variant in our dataset. While this version of replicon does occur outside Cluster A it is rare among CRhvKp (n=3/1,028) and CRKp (n=4/5,763) isolates, but more common among hvKp (n=184/988). The Cluster A variant has strong geographic and ST bias. It is found mostly in China (n=561/676), followed by Russia (n=22/676) and 18 other countries. KL64 (n=343/676) and KL1 (n=190/676) are the dominant serotypes, but these are likely the consequence of ST specificity being common in ST23 and ST11 types. Of all ST23 and ST11 isolates with any IncHI1B(pNDM-MAR) replicon, nearly all carried the Cluster A version of the replicon (ST23 197/216; ST11 461/525). In contrast, all isolates of the canonical hypervirulent ST86 (n=84/84) carried the PlasmidFinder's reference replicon variant.

While the IncHI1B(pNDM-MAR) replicon was nearly monophyletic in Cluster A (473 samples with identical replicon sequences and further 7 replicon sequences among 18 samples), across the entire dataset set IncHI1B(pNDM-MAR) replicon consisted of five main nucleotide sequences with only six mutation differences **(Data S2)**. Based on the context of Cluster A's *bla*$_{KPC-2}$ and IncHI1B(pNDM-MAR) replicon nucleotide sequence, there is little evidence that this cluster has generated an epidemic in countries outside of China that are well represented in our dataset. However, within China, Cluster A has been identified in at least 13 provinces since its first isolates were identified in 2013. The monitoring of the core replicon signatures of such Clusters can therefore assist with identifying the spread of CRhvKp forms.

DISCUSSION

In this study we examined plasmid replicons in a large global dataset of 12,468 Kp isolates with the aim of understanding global distribution of CRhvKp isolates. Our analysis revealed there is structure among the 1,028 CRhvKp isolates, with most belonging to multi-country clusters with at least nine clusters present globally. One such cluster (denoted as Cluster A) involved the spread of ST11 CRhvKp isolates and was detected by a surveillance system in China (17), and appears to have been contained in that country. The data also revealed the potential simultaneous spread of a smaller cluster in China involving ST15 strains with plasmid replicons identical to those found in smaller studies from Hangzou and Shanghai regions of China (21,40). Other clusters reveal outbreaks in multiple countries, where the exact dynamics of spread are harder to determine due to the limited geographic coverage.

We have focused on the plasmid replicons instead of sequence types (STs), because both hypervirulence associated and carbapenemase encoding genes are frequently found on plasmids and transfer horizontally. Hypervirulence associated genes exist almost exclusively on plasmids; while carbapenemase encoding genes are generally located on small insertion sequences or transposons embedded in plasmids (7,41). While STs are important in the context of outbreaks, we think that plasmid replicons provide greater insights into the spread of hypervirulence and carbapenemases, because plasmids are vectors that transfer AMR and virulence genes between bacterial strains. Due to relative stability of the replicon sequences, they can be used as signatures or barcodes of CRhvKp forms for clinical management, surveillance, and infection control activities.

Recent work on CRhvKp epidemiology using genomic data has focused on single hospitals or geographical regions (41,42). However, with increased human mobility and travel worldwide, large-scale analyses of all available sequences can provide insights from a global to local resolution. These activities can assist infection control decision making through identifying transmission hotspots, blind spots of sampling and informing resource allocation. Despite the global nature of our analysis, there are some geographical gaps due the convenience nature of the sampling, especially with lower numbers of samples from sub-Saharan Africa. While it is difficult to estimate the prevalence of plasmids and replicons, our analysis reveals the presence of their types and their diversity, and determines the spread of clusters. The number of actual circulating clusters is likely to be an underestimate and the dynamics of their spread incomplete. Large scale routine and timely sequencing globally can provide a more complete picture. This process should involve reviewing and updating the core clusters and their signatures, potentially using our statistical approach, which includes consideration of cluster robustness. Our statistical approach could be a robust alternative to the more general and widespread application of phylogenetic trees for cluster derivation, including bootstrapping for tree branch robustness, which may not be appropriate for Klebsiella and other pathogens with horizontal gene transfer (43,44).

Based on our analysis, heavily represented geographies with clusters of CRhvKp isolates (Thailand, Germany, Italy, USA, Russia) have not demonstrated rapid growth of a single cluster within a country. Instead, there were several small geographically diverse clusters, including one (Cluster H) which has a similar number of cases from Russia, Germany, and Egypt, whilst another (Cluster I) has isolates mainly from Italy, Russia, and China. Travel linked to tourism to an under-represented country is a plausible explanation. Relatedly, isolates collected across three Russian cities were present in six of nine CRhvKp clusters. This may be the linked to travel or importation events.

Another interesting finding was the presence of *K. aerogenes* isolates with salmochelin homologues. This *Klebsiella* species is not common in a clinical setting and generally has low level of AMR (38), but it may represent a potential reservoir of hypervirulent bacteria, especially in USA and Germany where it appears most common. Clinical isolates dominate most collections, but the collection and sequencing of samples from environmental or domestic animal sources may provide insights into AMR transmission. Relatedly, the comparison of the replicons from modern and Murray collection isolates showed perfectly matching replicon sequences in several cases often from isolates spanning 70 to 90 years. While there may be sample fidelity issues arising from a long-term storage, this may also be a result of a low mutation rate in replicon sequences. This in turn raises important questions about diversity of plasmids outside clinical environments. Insights into the diversity of plasmids, as well as the emergence and spread of AMR, will be provided through the future widespread application of cost-effective WGS or alternative amplicon sequencing-based approaches across countries and sampling sources combined with the development of informatic data platforms and large-scale analysis tools.

CONCLUSION

Our large-scale analysis of Kp isolates revealed nine global clusters of CRhvKp, including two expanding within China. We found multiple smaller multi-regional clusters that did not have a clear inter-temporal trend. The benefits of the surveillance and monitoring of infections using genomic data have been exemplified by global efforts during the recent COVID-19 pandemic. The generation of sequence data in real time can provide an early warning to infection control decision makers and provide clinical guidance to reduce burden. Relatedly, the presence of multi-country clusters reinforces that the sharing and joint analysis of international data provides important insights into the epidemiology of CRhvKp and other AMR infections, including the identification of countries experiencing outbreaks but potentially without extensive monitoring programs of their own. The goal would be to identify reservoirs of infections, assist the surveillance activities of infection control agencies and ultimately reduce the burden of Kp AMR.

DECLARATIONS

Ethics approval and consent to participate

Not required

Consent for publication

Not required

Availability of data and materials

The datasets supporting the conclusions of this article are available in the NCBI repository. Assemblies used are listed in **Data S1**. The data was visualised and quantified using a Python Notebook, which is available at https://github.com/AntonS-bio/KpReplicons/

Competing interests

The authors declare that they have no competing interests

Funding

TGC and SC are funded by the Medical Research Council UK grants (Reference no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1).

Authors' contributions

AS and TGC designed the study. AS analysed the data under the supervision of JP, SC and TGC. AS wrote the first draft of the manuscript, with contributions from JP, SC and TGC. All authors have edited manuscript drafts and agreed on the contents of the final version.

Acknowledgements

Not applicable

List of abbreviations

AMR – antimicrobial resistance

Kp – *Klebsiella pneumoniae*.

cKp – classic *K. pneumoniae*

hvKp – hypervirulent Kp defined as carrying either salmochelin or aerobactin locus, but not carbapenemase encoding genes.

CRKp – carbapenemase resistant Kp, defined as carbapenemase gene carrying Kp, but not salmochelin or aerobactin locus

CRhvKp - carbapenemase resistant hypervirulent Kp

MLST – multilocus sequence type

ST – sequence type

REFERENCES

1. Russo TA, Marr CM. Hypervirulent *Klebsiella pneumoniae* [Internet]. Vol. 32, Clinical Microbiology Reviews. American Society for Microbiology; 2019 [cited 2020 Sep 30]. Available from: https://pubmed.ncbi.nlm.nih.gov/31092506/

2. Zhang Y, Zhao C, Wang Q, Wang X, Chen H, Li H, et al. High prevalence of hypervirulent *Klebsiella pneumoniae* infection in China: Geographic distribution, clinical characteristics, and antimicrobial resistance. Antimicrob Agents Chemother [Internet]. 2016 Oct 1 [cited 2019 Dec 28];60(10):6115–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27480857

3. Struve C, Roe CC, Stegger M, Stahlhut SG, Hansen DS, Engelthaler DM, et al. Mapping the evolution of hypervirulent *Klebsiella pneumoniae*. mBio. 2015 Jul 21;6(4).

4. European Centre for Disease Prevention and Control. Risk Assessment: Emergence of hypervirulent *Klebsiella pneumoniae* ST23 carrying carbapenemase genes in EU/EEA countries. 2021.

5. Logan LK, Weinstein RA. The epidemiology of Carbapenem-resistant enterobacteriaceae: The impact and evolution of a global menace. Journal of Infectious Diseases [Internet]. 2017 [cited 2021 Mar 9];215(Suppl 1):S28–36. Available from: /pmc/articles/PMC5853342/

6. Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, et al. Clinical epidemiology of the global expansion of *Klebsiella pneumoniae* carbapenemases [Internet]. Vol. 13, The Lancet Infectious Diseases. NIH Public Access; 2013 [cited 2021 Mar 9]. p. 785–96. Available from: /pmc/articles/PMC4673667/

7. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. Vol. 31, Clinical Microbiology Reviews. American Society for Microbiology; 2018.

8. Partridge SR, Iredell JR. Genetic Contexts of blaNDM-1. Antimicrob Agents Chemother [Internet]. 2012 Nov [cited 2021 Oct 20];56(11):6065. Available from: /pmc/articles/PMC3486571/

9. Bush K, Bradford PA. Epidemiology of β-Lactamase-Producing Pathogens. Clin Microbiol Rev [Internet]. 2020 Apr 1 [cited 2022 Apr 7];33(2). Available from: https://pubmed.ncbi.nlm.nih.gov/32102899/

10. Russo TA, Olson R, Fang CT, Stoesser N, Miller M, MacDonald U, et al. Identification of Biomarkers for Differentiation of Hypervirulent *Klebsiella pneumoniae* from Classical K. pneumoniae. J Clin Microbiol [Internet]. 2018 Sep 1 [cited 2022 Jan 28];56(9). Available from: https://pubmed.ncbi.nlm.nih.gov/29925642/

11. Catalán-Nájera JC, Garza-Ramos U, Barrios-Camacho H. Hypervirulence and hypermucoviscosity: Two different but complementary *Klebsiella spp.* phenotypes? [Internet]. Vol. 8, Virulence. Taylor and Francis Inc.; 2017 [cited 2021 May 20]. p. 1111–23. Available from: /pmc/articles/PMC5711391/

12. Choby JE, Howard-Anderson J, Weiss DS. Hypervirulent *Klebsiella pneumoniae* – clinical and molecular perspectives. J Intern Med [Internet]. 2020 Mar 1 [cited 2021 Oct 11];287(3):283. Available from: /pmc/articles/PMC7057273/

13. Yao B, Xiao X, Wang F, Zhou L, Zhang X, Zhang J. Clinical and molecular characteristics of multi-clone carbapenem-resistant hypervirulent (hypermucoviscous) *Klebsiella pneumoniae* isolates in a tertiary hospital in Beijing, China. Int J Infect Dis [Internet]. 2015 Aug 1 [cited 2022 Apr 7];37:107–12. Available from: https://pubmed.ncbi.nlm.nih.gov/26141415/

14. 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 [Internet]. 2019 [cited 2019 Dec 28];15(4):e1008114. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30986243

15. Yang X, Sun Q, Li J, Jiang Y, Li Y, Lin J, et al. Molecular epidemiology of carbapenem-resistant hypervirulent *Klebsiella pneumoniae* in China. Emerg Microbes Infect [Internet]. 2022 Dec 31 [cited 2022 Apr 7];11(1):841–9. Available from: https://pubmed.ncbi.nlm.nih.gov/35236251/

16. Lam MMC, Wick RR, Wyres KL, Holt KE. Genomic surveillance framework and global population structure for *Klebsiella pneumoniae* [Internet]. bioRxiv. bioRxiv; 2020 [cited 2021 Mar 9]. p. 2020.12.14.422303. Available from: https://doi.org/10.1101/2020.12.14.422303

17. Liu C, Dong N, Chan EWC, Chen S, Zhang R. Molecular epidemiology of carbapenem-resistant *Klebsiella pneumoniae* in China, 2016-20. Lancet Infect Dis [Internet]. 2022 Feb 1 [cited 2022 Apr 7];22(2):167–8. Available from: https://pubmed.ncbi.nlm.nih.gov/35092791/

18. Yong M, Chen Y, Oo G, Chang KC, Chu WHW, Teo J, et al. Dominant Carbapenemase-Encoding Plasmids in Clinical Enterobacterales Isolates and Hypervirulent *Klebsiella pneumoniae*, Singapore. Emerg Infect Dis [Internet]. 2022 Aug 1 [cited 2022 Sep 15];28(8):1578–88. Available from: https://pubmed.ncbi.nlm.nih.gov/35876475/

19. Banerjee T, Wangkheimayum J, Sharma S, Kumar A, Bhattacharjee A. Extensively Drug-Resistant Hypervirulent *Klebsiella pneumoniae* From a Series of Neonatal Sepsis in a Tertiary Care Hospital, India. Front Med (Lausanne). 2021 Mar 8;8:186.

20. Kochan TJ, Nozick SH, Medernach RL, Cheung BH, Gatesy SWM, Lebrun-Corbin M, et al. Genomic surveillance for multidrug-resistant or hypervirulent *Klebsiella pneumoniae* among United States bloodstream isolates. BMC Infect Dis [Internet]. 2022 Dec 1 [cited 2022 Sep 15];22(1):1–21. Available from: https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-022-07558-1

21. Shu L, Dong N, Lu J, Zheng Z, Hu J, Zeng W, et al. Emergence of OXA-232 Carbapenemase-Producing *Klebsiella pneumoniae* That Carries a pLVPK-Like Virulence Plasmid among Elderly Patients in China. Antimicrob Agents Chemother [Internet]. 2019 Mar 1 [cited 2022 Sep 15];63(3). Available from: https://pubmed.ncbi.nlm.nih.gov/30559135/

22. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res [Internet]. 2016 [cited 2021 Jan 20];44(D1):D7–19. Available from: /pmc/articles/PMC4702911/?report=abstract

23. 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. Nature Communications 2021 12:1 [Internet]. 2021 Jul 7 [cited 2021 Oct 12];12(1):1–16. Available from: https://www.nature.com/articles/s41467-021-24448-3

24. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, et al. In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother [Internet]. 2014 [cited 2020 Nov 17];58(7):3895–903. Available from: https://pubmed.ncbi.nlm.nih.gov/24777092/

25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research [Internet]. 2011 [cited 2022 Mar 1];12(85):2825–30. Available from: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

26. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv [Internet]. 2018 Feb 9 [cited 2021 Apr 22]; Available from: http://arxiv.org/abs/1802.03426

27. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Vol. 9, Journal of Machine Learning Research. 2008.

28. Wickelmaier F. An Introduction to MDS. 2003;

29. Lin T, Zha H. Riemannian manifold learning. IEEE Trans Pattern Anal Mach Intell. 2008 May;30(5):796–809.

30. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet [Internet]. 2021 Jan 1 [cited 2022 Aug 4];66(1):85–91. Available from: https://pubmed.ncbi.nlm.nih.gov/33057159/

31. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol [Internet]. 2018 Jan 1 [cited 2022 Aug 4];37(1):38–47. Available from: https://pubmed.ncbi.nlm.nih.gov/30531897/

32. Yang Y, Sun H, Zhang Y, Zhang T, Gong J, Wei Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep [Internet]. 2021 Jul 27 [cited 2022 Aug 4];36(4). Available from: https://pubmed.ncbi.nlm.nih.gov/34320340/

33. Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. 2013 [cited 2022 Apr 11];7819 LNAI(PART 2):160–72. Available from: https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14

34. Baker KS, Burnett E, McGregor H, Deheer-Graham A, Boinett C, Langridge GC, et al. The Murray collection of pre-antibiotic era Enterobacteriacae: a unique research resource. Genome Med [Internet]. 2015 Dec 28 [cited 2020 Sep 23];7(1):97. Available from: http://genomemedicine.com/content/7/1/97

35. Bonnin RA, Jousset AB, Chiarelli A, Emeraud C, Glaser P, Naas T, et al. Emergence of new non-clonal group 258 high-risk clones among *Klebsiella pneumoniae* Carbapenemase-Producing K. Pneumoniae Isolates, France. Emerg Infect Dis [Internet]. 2020 Jun 1 [cited 2021 Feb 3];26(6):1212–20. Available from: https://doi.org/10.3201/eid2606.191517

36. Lam MMC, Wyres KL, Judd LM, Wick RR, Jenney A, Brisse S, et al. Tracking key virulence loci encoding aerobactin and salmochelin siderophore synthesis in *Klebsiella pneumoniae*. Genome Med [Internet]. 2018 Oct 29 [cited 2021 Mar 12];10(1). Available from: https://pubmed.ncbi.nlm.nih.gov/30371343/

37. Loraine J, Heinz E, de Sousa Almeida J, Milevskyy O, Voravuthikunchai SP, Srimanote P, et al. Complement Susceptibility in Relation to Genome Sequence of Recent *Klebsiella pneumoniae* Isolates from Thai Hospitals . mSphere [Internet]. 2018 Nov 7 [cited 2021 Jan 27];3(6). Available from: /pmc/articles/PMC6222052/?report=abstract

38. Passarelli-Araujo H, Palmeiro JK, Moharana KC, Pedrosa-Silva F, Dalla-Costa LM, Venancio TM. Genomic analysis unveils important aspects of population structure, virulence, and antimicrobial resistance in *Klebsiella aerogenes*. FEBS J [Internet]. 2019 Oct 1 [cited 2022 Apr 11];286(19):3797–810. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/febs.15005

39. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. Biol Direct. 2012;

40. Zhang Y, Chen C, Wu J, Jin J, Xu T, Zhou Y, et al. Sequence-Based Genomic Analysis Reveals Transmission of Antibiotic Resistance and Virulence among Carbapenemase-Producing *Klebsiella pneumoniae* Strains. mSphere [Internet]. 2022 Jun 29 [cited 2022 Sep 26];7(3). Available from: https://pubmed.ncbi.nlm.nih.gov/35546482/

41. Wyres KL, Nguyen TNT, Lam MMC, Judd LM, van Vinh Chau N, Dance DAB, et al. Genomic surveillance for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from South and Southeast Asia. Genome Med [Internet]. 2020 Jan 16 [cited 2021 Jan 22];12(1). Available from: /pmc/articles/PMC6966826/?report=abstract

42. Hawkey J, Wyres KL, Judd LM, Harshegyi T, Blakeway L, Wick RR, et al. ESBL plasmids in *Klebsiella pneumoniae*: diversity, transmission and contribution to infection burden in the

hospital setting. Genome Med [Internet]. 2022 Dec 1 [cited 2022 Sep 26];14(1):97. Available from: https://pubmed.ncbi.nlm.nih.gov/35999578/

43. Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. Elife. 2021 Jan 1;10:1–61.

44. Stott CM, Bobay LM. Impact of homologous recombination on core genome phylogenies. BMC Genomics [Internet]. 2020 Dec 1 [cited 2021 Jun 10];21(1):1–10. Available from: https://doi.org/10.1186/s12864-020-07262-x

TABLES

Table 1

List of CRhvKp clusters. In some assemblies carbapenemase genes (CR), *iuc* and *iro* (hypervirulence; HV) loci were located on the same assembly contig or chromosome. The underlying data is available in Supplementary Data S2, S3 and S4.

Cluster	Is robust*	N	Main STs (N)	Total STs	Species	Collection Dates	Main Replicons**	Main Countries	Main Siderophores	Main Carbapenemases	Contig with CR and replicon	Contig with HV and replicon	Contig with CR and HV
A	Yes (0%)	560	ST11(499), ST23(35), ST268(13)	11	*K. pneumoniae,* *K. quasipneumoniae*	2012-2021	ColRNAI(943), repB(555), IncFII(pHN7A8)(501), IncHI1B(pNDM-MAR)(489), IncR(486)	China(517), Russia(12), Missing(8), Singapore(6)	iuc 1(554), iro 1(90), iro 1 (incomplete) (truncated)(22), iro 1 (incomplete)(55)	OXA-48(16), KPC-2(516), IMP-4(8)	74	341	2
B	No (99%)	83	ST11(74)	5	*K. pneumoniae*	2012-2021	ColRNAI(139), IncFII(pHN7A8)(74), IncR(60)	China(81)	iuc 1(78)	KPC-2(79)	20	11	0
C	Yes (0%)	82	ST147(32), ST395(24), ST383(10)	15	*K. aerogenes,* *K. pneumoniae*	2012-2020	Col(pHAD28)(91), IncHI1B(pNDM-MAR)(80), IncFIB(pNDM-Mar)(79)	Russia(32), China(16), Egypt(9), Germany(8)	iuc 1(76)	OXA-48(51), NDM-1(23)	16	11	3
D	Yes (0%)	61	ST231(60)	2	*K. pneumoniae*	2013-2019	Col440I(73), IncFIB(pQil)(61), IncFII(K)(61), IncFIA(60), IncFII(pAMA1167-NDM-5)(60), ColKP3(59), Col(pHAD28)(55)	Thailand(24), India(17), Pakistan(8)	iuc 5(60)	OXA-232(59)	60	8	0
E	No (88%)	50	ST86(16), ST65(14)	10	*K. pneumoniae*	2013-2019	repB(50), IncHI1B(pNDM-MAR)(49)	China(16), Singapore(14), Thailand(13)	iuc 1(44), iro 1(40)	KPC-2(32), OXA-232(13)	19	32	0
F	Yes (0%)	34	ST290(7), ST15(6)	12	*K. pneumoniae*	2012-2020	IncFII(K)(35), IncFIB(K)(32)	China(23), Russia(7)	iuc 3 (truncated)(16), iuc 1(14)	NDM-5(8), NDM-1(14), KPC-2(8)	8	19	2
G	Yes (0%)	32	ST15(26)	6	*K. pneumoniae*	2015-2020	repB(31), ColRNAI(31), IncHI1B(pNDM-MAR)(30), IncFIB(pKPHS1)(28), IncFII(K)(28), ColKP3(28), Col(pHAD28)(28), Col440I(25), IncFIB(pQil)(21)	China(25)	iuc 1(32)	OXA-232(27)	32	7	0
H	Yes (0%)	31	ST512(9)	6	*K. aerogenes,* *K. pneumoniae*	2008-2021	IncFII(K)(23), ColRNAI(22), IncFIB(pQil)(20)	Italy(11), USA(6)	iuc 1(16), iro unknown(13)	KPC-3(15)	18	2	0
I	Yes (0%)	16	ST395(16)	1	*K. pneumoniae*	2015-2017	IncHI1B(pNDM-MAR)(16), IncFIB(K)(16), Col440II(16), Col(pHAD28)(16), IncFII(K)(16), IncFIB(pNDM-Mar)(15)	Italy(16)	iuc 1(15)	KPC-3(14)	2	5	0
Unassigned	Yes (0%)	79	ST101(8), ST2096(6), ST218(6)	25	*K. aerogenes, K. variicola subsp. variicola, K. pneumoniae*	2011-2021	Col(pHAD28)(45), IncFII(K)(40), ColRNAI(38), IncHI1B(pNDM-MAR)(30), IncFIB(pNDM-Mar)(28)	China(16), USA(11), Germany(8), Russia(7)	iuc 1(38), iuc 5(7), iro unknown(29), iro 1(11)	KPC-3(9), KPC-2(19), OXA-48(13), OXA-232(8), OXA-244(6)	32	17	1

HV - hypervirulent Kp; CRKp – carbapenemase resistant Kp; * Cluster is considered robust if <1% of -$\log_{10}$(chi-squared p-value) are <20, see Methods for detail ** Different clusters with the same main replicons may have different variants.

FIGURES

Figure 1

Frequency of the main nucleotide sequences from the IncFIB(K) replicon family. The eight nucleotide sequences are provided in Data S2.

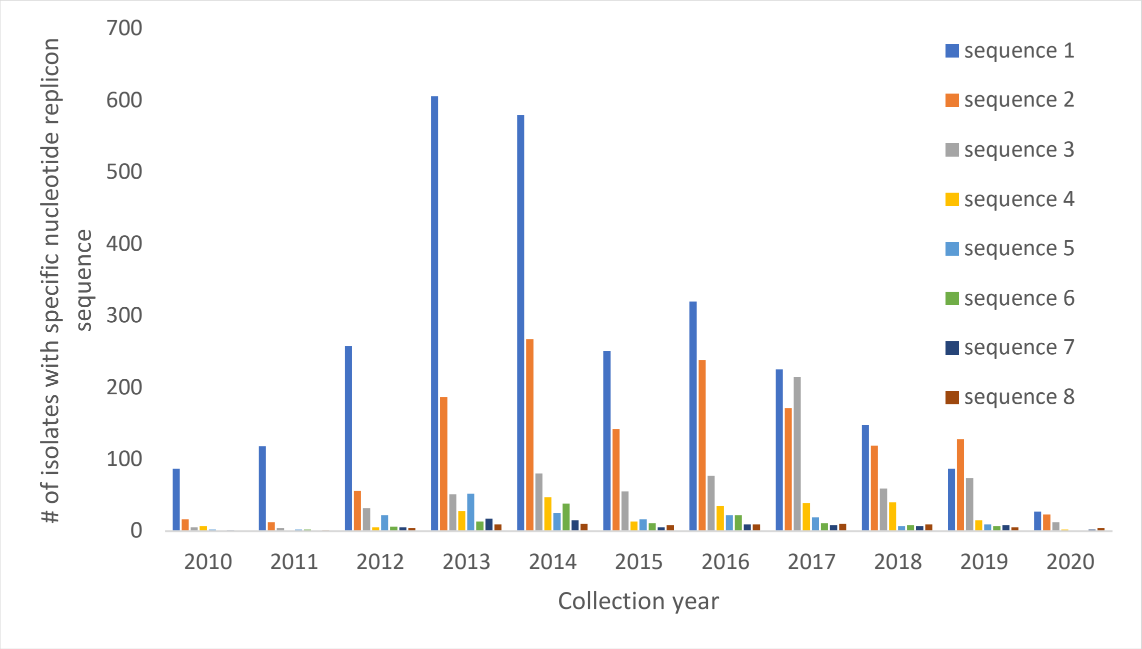


Figure 2

Clustering of CRKp, hvKp, and CRhvKp Kp isolates based on replicon sequences coloured by (A) major sequence types (STs), (B) country, and (C) genotype. In (D) non-CRhvKp isolates are hidden and only CRhvKp isolates are visible and coloured by plasmid cluster. Axes are dimensionless. Underlying data is presented in Data S3.

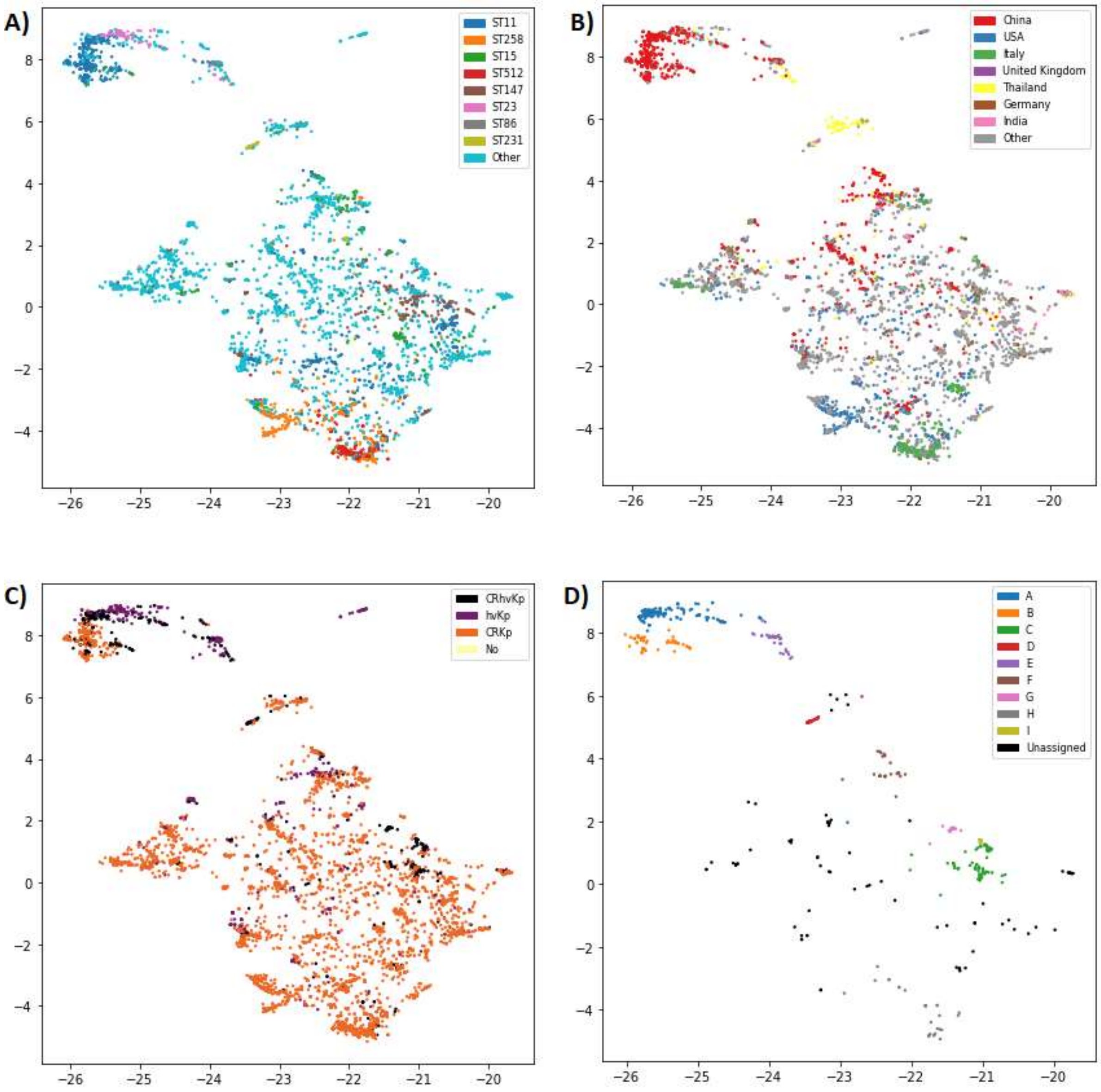
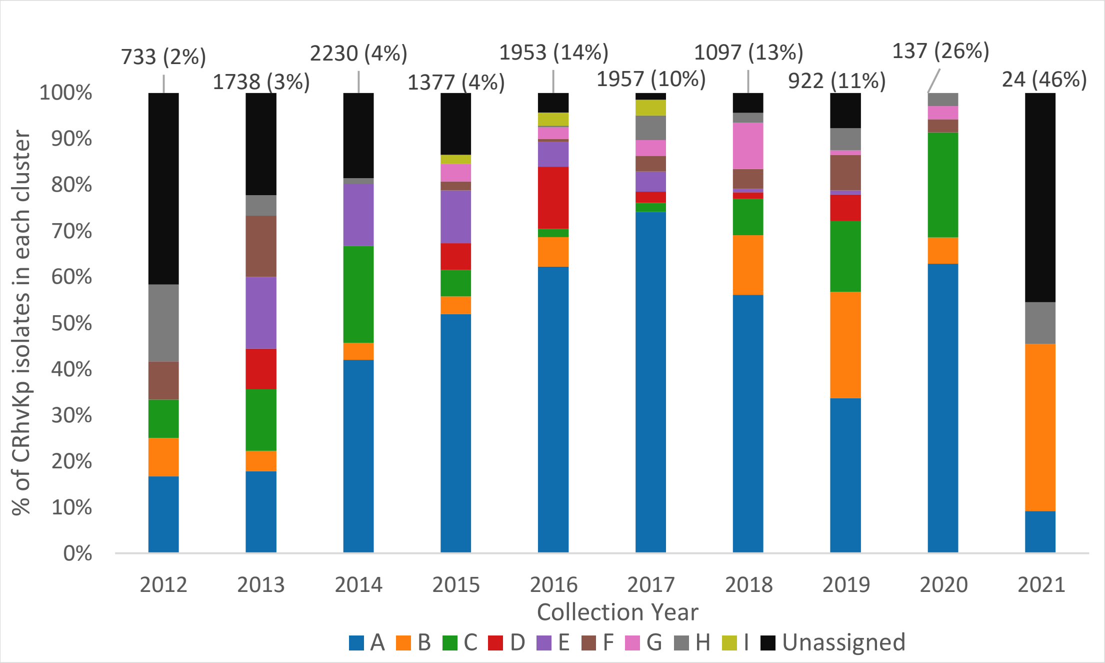


Figure 3

Relative abundance and geographic diversity of carbapenem resistant hypervirulent Kp (CRhvKp) isolates. Total number of all genotyped isolates in each year and percentage which are CRhvKp at the top.



Chapter 7: Discussion

Kp genomics can assist with understanding how the bacterium transmits, causes infections and disease, and resists treatments. For example, understanding of Kp methylome and whether it has a role in limiting dissemination of plasmids between cells will help predict the spread of AMR genes within different geographic regions. In a more direct way, the identification of virulence genes may not only assist in the epidemiological monitoring of hypervirulent strains, but also provide a list of proteins that can be targeted with therapy. Furthermore, understanding Kp genomics improves the understanding of the AMR drivers. While levels of AMR are routinely monitored in many countries, the underlying causes are much less explored. Understanding the mode of resistance can not only guide control of AMR, but also, together with population structure analysis, give real-time feedback on efficacy of AMR control methods.

In the study of methylation in Kp, I analysed PacBio sequencing data from eight Portuguese Kp isolates, which had undergone antibiotic susceptibility phenotyping. I used this data to characterise the bacterial epigenome and explore the relationship between methylation and AMR. I focused on methylation (including around AMR genes) and on differences in the abundance of R-M recognition motifs on Kp chromosomes and MGEs. The abundance of some target methylation motifs was different between chromosomes and plasmids, especially the GATC motifs methylated by orphan MTase Dam. I also found that a GATC motif immediately downstream of the *fosA* gene, which confers low level fosfomycin resistance, was consistently unmethylated in the samples. Isolates that had the *tnpB* transposase gene on the IncFIB(K) plasmid also consistently lacked methylation immediately downstream of this gene.

Due to a small sample size, these results are difficult to generalise. I found no statistically significant difference between the frequency of non-GATC R-M recognition motifs on plasmids and chromosomes either within or between examined strains. However, this maybe the result of small samples size or sample selection. The ideal dataset would be substantially larger and consist of two sets of isolates: isolates from the same ST and isolates from diverse STs. Together, these would give understanding of the diversity of R-M systems and the role of recognition motifs in shaping the Kp genome. The isolates from the same ST would elucidate if R-M systems persist in the bacterial lineages. The comparison between STs would reveal if isolates' genomes are enriched for motifs recognised by R-M systems carried by the isolates. As the R-M systems are frequently carried on plasmids, this information would give insights into plasmid persistence within bacterial strains.

In the study of longitudinal isolates from Portugal I aimed to improve the understanding of the genomic landscape of Kp in that country. To do that, I analysed 509 WGS isolates spanning a period from years 1980 to 2019 - the largest dataset to date. I found that although sequence types ST15, ST14 and ST147 predominate, almost one-third of isolates came from STs considered infrequent in Portugal. I established that there are many AMR determinants and, as expected, these evolved over time. In the 1980's, the broad spectrum beta-lactamases encoded by *bla*$_{OXA\text{-}9}$ and *bla*$_{TEM\text{-}1}$ genes were prevalent, until the appearance and rapid spread of ESBL *bla*$_{CTX\text{-}M\text{-}15}$ in early 2000's, which in turn gave way to the carbapenemase gene *bla*$_{KPC\text{-}3}$ since 2010's. The analysis of the underlying plasmids revealed a mosaic distribution across isolates and time, suggestive of strong selective pressures. This work provides a baseline set of variants for future AMR monitoring and epidemiological studies in Portugal and wider Europe. This work also highlighted the correlation between plasmid replicons and AMR genes. Plasmids, rather than STs, appear better predictors of the AMR gene population structure.

While the study examines a very large dataset, it does not give insight into the diversity of Kp in environmental and asymptomatic community reservoirs. The strength of this dataset is that it permits identification of discrepancies between phenotype and genotype, which in turn allows identification of AMR drivers. Using this dataset, I found that *bla*$_{SHV\text{-}28}$ (class A beta-lactamase) has an unexpected ability to inhibit cefoxitin, which is a phenotype normally associated with class C beta-lactamases that were rare in this dataset. Similar insights may be gained through a national program of routine sequencing and phenotyping of Kp and other pathogens. For example, isolates carrying only *bla*$_{KPC}$ and resistant to combination of beta-lactam antibiotics and beta-lactamase inhibitors are currently rare. A national program aimed at reconciling AMR phenotype and genotype would enable early detection of such isolates and would help limit their spread.

Another limitation of the dataset used in **Chapter 4** is the national sampling which results the lack of depth at the local level. This lack of depth limits the ability to identify reservoirs of AMR genes and their transmission routes into hospital settings. For example, the lack of environmental and community samples meant I was unable to determine if hospitals themselves are reservoirs of AMR genes, or if AMR genes regularly enter the hospital system from outside. This question is very important from a public health perspective because the answer will dictate where to target infection control measures. The best dataset to answer this question would select a small geographic region and collect samples from both clinical and environmental sources over a period of time. Such a dense local dataset can be combined with national monitoring data. Together these datasets would allow identification of local transmission dynamics and importation of strains from outside the densely sampled region.

In the study of genes linked to the hypervirulent phenotype of Kp, I analysed the shared accessory genome of all publicly available Kp samples sourced from the liver (n=79) and compared them to a large globally diverse public Kp dataset (n=646). My aim was to determine what genes are associated with the liver invasive phenotype. Consistent with previous research (15,189,190), I found links to both *iuc* and *iro* loci. The results suggest that these siderophore systems likely belong to different MGEs. In line with previous results (91), the hypermucoidy associated genes *rmpA* and *rmpC* were not linked to hypervirulence. I have found that in hypervirulent phenotypes, *iuc* and *iro* loci are consistently accompanied by an additional 12 and 13 genes, respectively. This insight suggests a possible functional link that could be exploited for the development of diagnostics and therapies. My approach found no other genes or SNVs that had a similarly strong link to a liver invasive phenotype. While *iuc* and maybe *iro* loci are necessary for the hypervirulence phenotype, they alone may be insufficient. As a part of this study, I also determined that hypervirulence plasmids in Kp are likely composed of blocks of genes with genes in individual blocks usually being absent or present simultaneously. The unresolved question of this study is the role of additional (non-*iuc/iro*) genes in Kp hypervirulence. This question can only be answered through experimental studies. In addition, a larger dataset of Kp with hypervirulent clinical presentation may help identify additional genes linked to hypervirulence.

I applied the insights from the study of AMR genes in Portuguese dataset and the hypervirulence genes to a global dataset (n=13,176) to understand the geographical dynamics of hypervirulence and carbapenemase genes, as well as plasmid replicons. I examined plasmid replicons to understand the global distribution of carbapenem resistant hypervirulent (CRhvKp) isolates. This analysis revealed there is structure among the ~1k CRhvKp isolates, with most belonging to multi-country clusters, with at least nine present globally. One such cluster involved the spread of ST11 CRhvKp isolates previously identified by a surveillance system in China (191), but the data also revealed the simultaneous spread of a smaller cluster in the same country involving ST15 strains. Despite rapid spread, the largest cluster appeared contained in the country. Other clusters reveal multiple cross-country outbreaks unable to spread. However, an alternative explanation for small clusters is that I detected only echoes of the larger outbreaks because the outbreaks are present in poorly sampled countries. Thus, an improvement on this study would be additional sampling of underrepresented countries and inclusion of samples that have been sequenced, but not assembled. The latter could triple the number of available isolates.

My work can be extended by the development of a website that shows a global distribution hypervirulence biomarkers in Kp and clusters of CRhvKp. Making this information readily available in a user-friendly format would assist infection control and decision making. The increasing availability of sequencing technologies with low fixed cost such as MinIon or Flongle from ONT should increase the global sequencing coverage, including within Africa. Together, these two tools would allow detection of rapidly spreading CRhvKp variants at the stage when their overall prevalence remains low.

In general, the examination of available data revealed that some geographic regions are strongly overrepresented while some have very limited data. This bias means that the true diversity of Kp is unknown. To put this in context, China and Portugal both have enough data to understand the dominant AMR and virulence genotypes. This data demonstrates that in both countries *bla_{KPC}* is the dominant carbapenemase encoding gene, but the alleles (*bla_{KPC-3}* versus *bla_{KPC-2}*) and genetic contexts (Tn4401 versus NTE_{KPC}) are completely different (192). Furthermore, in Portugal, unlike China, the hypervirulent genotype is rare. Such differences exist between many other countries; thus, knowledge of AMR and virulence drivers present in one country may not apply to another. Understanding the diversity of Kp, or any other pathogen, helps inform effective public health policy. Even in a high resource setting, the clinical decision making is usually driven by phenotype focused methods such as disc diffusion. The comprehensive sequencing of isolates is also rare and mostly focuses on nationally relatively rare infections, such as *M. tuberculosis* in the United Kingdom. However, even a limited national sequencing program can help guide policy decisions, especially in low resource settings. For example, knowledge of nationally dominant beta-lactamases can guide procurement of beta-lactamase inhibitors; while understanding of virulence profiles can help shape clinical guidance when doctors are faced with symptoms of hypervirulent Kp.

I have focused on the plasmid replicons instead of STs because both hypervirulence associated and carbapenemase encoding genes are frequently found on plasmids and transfer horizontally. Hypervirulence associated genes exist almost exclusively on plasmids, while carbapenemase encoding genes are generally located on small ISs or transposons embedded in plasmids (94,193). While STs are undoubtedly important in the context of outbreaks, my analysis reveals that fine-scale dynamics of population structure are informed by plasmid replicons. The use of plasmid replicons does not disregard STs. When a particular combination of plasmids is spread clonally, the corresponding isolates will form a cluster based on replicons. An example of this is a cluster located at the intersection of ST23 and a subset of ST11, found in **Chapter 6**. ST23 is canonical hypervirulent strain while ST11 is dominant carbapenem resistant strain in China. These two STs are very distinct with ST11 being closer to ST258, the dominant strain in the USA (194). A phylogeny-free approach can also be used to search for a ST11-ST23 hybrid strain which, to my knowledge, has not been reported yet.

The limitation of my plasmid focused analysis is that based on previous findings (195), which my work replicated, the majority of Kp genes are accessory. Conjugation and transduction are common in Kp, and account for some of the movement of accessory genes between strains (19,106). However, the other common mechanism, transformation, probably has a limited role because Kp is not known to be naturally competent (94,95). Understanding the main route by which Kp isolates can acquire chromosomal DNA would substantially improve our understanding of Kp epidemiology. At present, there is limited understanding of the frequency and mode of horizontal chromosomal genes transfer in Kp, and there is debate about frequency of importation of genes to chromosomes (18,163,196). This has a direct impact on outbreak investigations and population genomics. If horizontal gene transfer is limited, it can be accounted for in phylogenetic trees (164,165). However, if the nucleotide diversity is driven by exchange of SNPs via horizontal gene transfer, this diminishes the utility of phylogenetic trees in outbreak and epidemiological investigations (18). While bioinformatic approaches can help identify the likely presence of recombination, most approaches (164,165) focus on detection of infrequent recombination between clones that have accumulated sufficient number of distinguishing mutations. Experimental work on crossing strains combined with long-read sequencing can help better understand the frequency of recombination and thus inform the appropriate methods for epidemiological inference.

Chapter 8: Conclusion

This thesis has examined various aspects of the non-core genome of Kp, especially those linked to plasmids. I examined the methylation patterns and linked genes using a set of clinical Kp samples from Portugal. I used a larger set of Kp isolates to understand the temporal evolution of AMR profiles in Portugal. This work pointed to plasmids as the determinant of the beta-lactamase AMR profile of Kp. I have examined a dataset of phenotypically hypervirulent Kp to determine the biomarkers of this phenotype. The identified biomarkers included known siderophores, but also additional genes that frequently accompany siderophores on the hypervirulence plasmids. In the subsequent analysis, I unified these themes by examining the plasmids of genotypic hvKp and carbapenem resistant Kp isolates. The results revealed that plasmids can explain the origin of CRhvKp genotypes and assist the identification of clusters of related cases, which may span multiple countries. With whole genome sequencing of Kp gaining traction globally, my work reinforces the benefits of using the resulting data to inform on AMR and hypervirulence mechanisms, understand transmission patterns, and develop much needed interventions and informatic tools for infection control.

References

1. Bagley ST. Habitat Association of *Klebsiella* Species. Infect Control Hosp Epidemiol. 1985;6(2):52–8.
2. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018 Jan 1;46(D1):D8–13.
3. Neog N, Phukan U, Puzari M, Sharma M, Chetia P. *Klebsiella oxytoca* and Emerging Nosocomial Infections. Current Microbiology 2021 78:4. 2021 Mar 3;78(4):1115–23.
4. Rodríguez-Medina N, Barrios-Camacho H, Duran-Bedolla J, Garza-Ramos U. *Klebsiella variicola*: an emerging pathogen in humans. Emerg Microbes Infect. 2019 Jan 1;8(1):973.
5. Wesevich A, Sutton G, Ruffin F, Park L, Fouts D, Fowler V, et al. Newly Named *Klebsiella aerogenes* (formerly *Enterobacter aerogenes*) Is Associated with Poor Clinical Outcomes Relative to Other *Enterobacter* Species in Patients with Bloodstream Infection. J Clin Microbiol. 2020 Sep 1;58(9).
6. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. Curr Opin Genet Dev. 2005 Dec 1;15(6):589–94.
7. Wyres KL, Holt KE. *Klebsiella pneumoniae* Population Genomics and Antimicrobial-Resistant Clones. Trends Microbiol. 2016 Dec 1;24(12):944–56.

8. Yigit H, Queenan A, Anderson G, Domenech-Sanchez A, Biddle J, Steward C, et al. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of *Klebsiella pneumoniae*. Antimicrob Agents Chemother. 2001;45(4):1151–61.

9. Findlay J, Poirel L, Kessler J, Kronenberg A, Nordmann P. New Delhi Metallo-β-Lactamase-Producing *Enterobacterales* Bacteria, Switzerland, 2019-2020. Emerg Infect Dis. 2021 Oct 1;27(10):2628–37.

10. Porreca AM, Sullivan K v., Gallagher JC. The Epidemiology, Evolution, and Treatment of KPC-Producing Organisms. Current Infectious Disease Reports 2018 20:6. 2018 May 5;20(6):1–12.

11. Duin D van, Doi Y. The global epidemiology of carbapenemase-producing *Enterobacteriaceae*. Virulence. 2017 May 19;8(4):460.

12. WHO. Antimicrobial resistance. Global report on surveillance. World Health Organization [Internet]. 2014 [cited 2021 Oct 11];61(3):12–28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22247201%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2536104&tool=pmcentrez&rendertype=abstract

13. Diancourt L, Passet V, Verhoef J, Grimont PAD, Brisse S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. J Clin Microbiol. 2005 Aug;43(8):4178–82.

14. Zhou H, Liu W, Qin T, Liu C, Ren H. Defining and Evaluating a Core Genome Multilocus Sequence Typing Scheme for Whole-Genome Sequence-Based Typing of *Klebsiella pneumoniae*. Front Microbiol. 2017;8:371.

15. Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard AS, et al. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. Emerg Infect Dis. 2014 Nov 1;20(11):1812–20.

16. 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.

17. Comandatore F, Sassera D, Bayliss SC, Scaltriti E, Gaiarsa S, Cao X, et al. Gene Composition as a Potential Barrier to Large Recombinations in the Bacterial Pathogen *Klebsiella pneumoniae*. Whitaker R, editor. Genome Biol Evol. 2019 Nov 1;11(11):3240–51.

18. Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. Elife. 2021 Jan 1;10:1–61.

19. Wyres KL, Holt KE. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. Curr Opin Microbiol. 2018 Oct 1;45:131–9.

20. Yuval B, Ben-Ami E, Behar A, Ben-Yosef M, Jurkevitch E. The Mediterranean fruit fly and its bacteria – potential for improving sterile insect technique operations. Journal of Applied Entomology. 2013 Jun;137(SUPPL.1):39–42.

21. Conlan S, Kong HH, Segre JA. Species-Level Analysis of DNA Sequence Data from the NIH Human Microbiome Project. PLoS One. 2012 Oct 10;7(10).

22. Gorrie CL, Mirc Eta M, Wick RR, Edwards DJ, Thomson NR, Strugnell RA, et al. Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients. Clinical Infectious Diseases. 2017 Jul 15;65(2):208–15.

23. Choby JE, Howard-Anderson J, Weiss DS. Hypervirulent *Klebsiella pneumoniae* – clinical and molecular perspectives. J Intern Med. 2020 Mar 1;287(3):283.

24. Dao TT, Liebenthal D, Tran TK, Vu BNT, Nguyen DNT, Tran HKT, et al. *Klebsiella pneumoniae* Oropharyngeal Carriage in Rural and Urban Vietnam and the Effect of Alcohol Consumption. PLoS One. 2014 Mar 25;9(3).

25. Carpenter J. *Klebsiella* pulmonary infections: occurrence at one medical center and review. Rev Infect Dis. 1990;12(4):672–82.

26. Ko WC, Paterson DL, Sagnimeni AJ, Hansen DS, Gottberg A von, Mohapatra S, et al. Community-Acquired *Klebsiella pneumoniae* Bacteremia: Global Differences in Clinical Patterns. Emerg Infect Dis. 2002;8(2):160.

27. Clegg S, Murphy CN. Epidemiology and Virulence of *Klebsiella pneumoniae*. Microbiol Spectr. 2016 Feb 2;4(1).

28. Jones RN, Flonta M, Gurler N, Cepparulo M, Mendes RE, Castanheira M. Resistance surveillance program report for selected European nations (2011). Diagn Microbiol Infect Dis. 2014 Apr 1;78(4):429–36.

29. European Centre for Disease Prevention and Control. Antimicrobial resistance in the EU/EEA (EARS-Net) - Annual Epidemiological Report for 2019 [Internet]. 2020 [cited 2021 Mar 9]. Available from: https://www.ecdc.europa.eu/en/publications-data/surveillance-antimicrobial-resistance-europe-2019

30. Centers for Disease Control and Prevention. ANTIBIOTIC RESISTANCE THREATS IN THE UNITED STATES. 2019.

31. Ramirez MS, Tolmasky ME. Aminoglycoside modifying enzymes. Drug Resistance Updates. 2010;13(6):151–71.

32. Redgrave LS, Sutton SB, Webber MA, Piddock LJV. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. Trends Microbiol. 2014 Aug 1;22(8):438–45.

33. Bonomo RA. β-Lactamases: A focus on current challenges. Vol. 7, Cold Spring Harbor Perspectives in Medicine. Cold Spring Harbor Laboratory Press; 2017.

34. Wang R, Dorp L van, Shaw LP, Bradley P, Wang Q, Wang X, et al. The global distribution and spread of the mobilized colistin resistance gene *mcr-1*. Nat Commun. 2018 Dec 1;9(1).

35. Drawz SM, Bonomo RA. Three Decades of β-Lactamase Inhibitors. Clin Microbiol Rev. 2010 Jan;23(1):160.

36. Papp-Wallace KM, Endimiani A, Taracila MA, Bonomo RA. Carbapenems: Past, Present, and Future. Antimicrob Agents Chemother. 2011 Nov;55(11):4943.

37. Varela MF, Stephen J, Lekshmi M, Ojha M, Wenzel N, Sanford LM, et al. Bacterial Resistance to Antimicrobial Agents. Antibiotics. 2021;10(5).

38. Cross T, Ransegnola B, Shin JH, Weaver A, Fauntleroy K, VanNieuwenhze MS, et al. Spheroplast-Mediated Carbapenem Tolerance in Gram-Negative Pathogens. Antimicrob Agents Chemother. 2019;63(9).

39. Jacoby GA, Mills DM, Chow N. Role of β-Lactamases and Porins in Resistance to Ertapenem and Other β-Lactams in *Klebsiella pneumoniae*. Antimicrob Agents Chemother. 2004 Aug;48(8):3203.

40. Bush K, Jacoby GA, Medeiros AA. A functional classification scheme for beta-lactamases and its correlation with molecular structure. Antimicrob Agents Chemother. 1995;39(6):1211.

41. Ambler RP, Coulson AF, Frère JM, Ghuysen JM, Joris B, Forsman M, et al. A standard numbering scheme for the class A beta-lactamases. Biochemical Journal. 1991;276(Pt 1):269.

42. Carcione D, Siracusa C, Sulejmani A, Leoni V, Intra J. Old and New Beta-Lactamase Inhibitors: Molecular Structure, Mechanism of Action, and Clinical Use. Antibiotics. 2021 Aug 1;10(8).

43. Tooke CL, Hinchliffe P, Bragginton EC, Colenso CK, Hirvonen VHA, Takebayashi Y, et al. β-Lactamases and β-Lactamase Inhibitors in the 21st Century. J Mol Biol. 2019 Aug 23;431(18):3472.

44. Räisänen K, Koivula I, Ilmavirta H, Puranen S, Kallonen T, Lyytikäinen O, et al. Emergence of ceftazidime-avibactam-resistant *Klebsiella pneumoniae* during treatment, Finland, December 2018. Eurosurveillance. 2019;24(19):1.

45. Mojica MF, Rossi MA, Vila AJ, Bonomo RA. The urgent need for metallo-β-lactamase inhibitors: an unattended global threat. Lancet Infect Dis. 2021 Jul;0(0).

46. Lasko MJ, Nicolau DP. Carbapenem-Resistant *Enterobacterales*: Considerations for Treatment in the Era of New Antimicrobials and Evolving Enzymology. Curr Infect Dis Rep. 2020 Mar 1;22(3).

47. Kim SW, Lee JS, Park S bin, Lee AR, Jung JW, Chun JH, et al. The Importance of Porins and β-Lactamase in Outer Membrane Vesicles on the Hydrolysis of β-Lactam Antibiotics. Int J Mol Sci. 2020 Apr 2;21(8).

48. Dulyayangkul P, Ismah WAKWN, Douglas EJA, Avison MB. Mutation of *kvrA* Causes OmpK35 and OmpK36 Porin Downregulation and Reduced Meropenem-Vaborbactam Susceptibility in KPC-Producing *Klebsiella pneumoniae*. Antimicrob Agents Chemother. 2020 Jul 1;64(7).

49. Albiger B, Glasner C, Struelens MJ, Grundmann H, Monnet DL, group the ES of CPE (EuSCAPE) working. Carbapenemase-producing *Enterobacteriaceae* in Europe: assessment by national experts from 38 countries, May 2015. Eurosurveillance. 2015 Nov 12;20(45):30062.

50. Logan LK, Weinstein RA. The epidemiology of Carbapenem-resistant *Enterobacteriaceae*: The impact and evolution of a global menace. Journal of Infectious Diseases. 2017;215(Suppl 1):S28–36.

51. Bonomo RA, Burd EM, Conly J, Limbago BM, Poirel L, Segre JA, et al. Carbapenemase-Producing Organisms: A Global Scourge. Clin Infect Dis. 2018 Apr 3;66(8):1290.

52. Boyd SE, Livermore DM, Hooper DC, Hope WW. Metallo-β-lactamases: Structure, function, epidemiology, treatment options, and the development pipeline. Antimicrob Agents Chemother. 2020 Oct 1;64(10).

53. Zárate SG, Claure MLD la C, Benito-Arenas R, Revuelta J, Santana AG, Bastida A. Overcoming Aminoglycoside Enzymatic Resistance: Design of Novel Antibiotics and Inhibitors. Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry. 2018;23(2).

54. Cox G, Ejim L, Stogios PJ, Koteva K, Bordeleau E, Evdokimova E, et al. Plazomicin Retains Antibiotic Activity against Most Aminoglycoside Modifying Enzymes. ACS Infect Dis. 2018 Jun 8;4(6):980.

55. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2020 Jan 1;48(D1):D517.

56. Saravolatz LD, Stein GE. Plazomicin: A New Aminoglycoside. Clinical Infectious Diseases. 2020 Feb 3;70(4):704–9.

57. Redgrave LS, Sutton SB, Webber MA, Piddock LJV. Fluoroquinolone resistance: Mechanisms, impact on bacteria, and role in evolutionary success. Vol. 22, Trends in Microbiology. Elsevier Ltd; 2014. p. 438–45.

58. Lindgren PK, Marcusson LL, Sandvang D, Frimodt-Møller N, Hughes D. Biological Cost of Single and Multiple Norfloxacin Resistance Mutations in *Escherichia coli* Implicated in Urinary Tract Infections. Antimicrob Agents Chemother. 2005 Jun;49(6):2343.

59. Morgan-Linnell SK, Zechiedrich L. Contributions of the Combined Effects of Topoisomerase Mutations toward Fluoroquinolone Resistance in *Escherichia coli*. Antimicrob Agents Chemother. 2007 Nov;51(11):4205.

60. Robicsek A, Strahilevitz J, Jacoby G, Macielag M, Abbanat D, Park C, et al. Fluoroquinolone-modifying enzyme: a new adaptation of a common aminoglycoside acetyltransferase. Nat Med. 2006 Jan;12(1):83–8.

61. Aghamali M, Sedighi M, Zahedi A, Mohammadzadeh N, Abbasian S, Ghafouri Z, et al. Fosfomycin: mechanisms and the increasing prevalence of resistance. Journal of Medical Biology. 2019;68.

62. Karageorgopoulos DE, Wang R, Yu XH, Falagas ME. Fosfomycin: evaluation of the published evidence on the emergence of antimicrobial resistance in Gram-negative pathogens. J Antimicrob Chemother. 2012 Feb;67(2):255–68.

63. Ito R, Mustapha MM, Tomich AD, Callaghan JD, McElheny CL, Mettus RT, et al. Widespread fosfomycin resistance in gram-negative bacteria attributable to the chromosomal *fosA* gene. mBio. 2017 Jul 1;8(4).

64. Liu YY, Wang Y, Walsh TR, Yi LX, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis. 2016 Feb 1;16(2):161–8.

65. Hubert EG, Potter CS, Hensley TJ, Cohen M, Kalmanson GM, Guze LB. L-Forms of *Pseudomonas aeruginosa*. Infect Immun. 1971;4(1):60.

66. EUCAST: Clinical breakpoints and dosing of antibiotics - v 11.0 [Internet]. 2021 [cited 2021 Mar 9]. Available from: https://www.eucast.org/clinical_breakpoints/

67. Grossman TH. Tetracycline antibiotics and resistance. Cold Spring Harb Perspect Med. 2016 Apr 1;6(4):a025387.

68. Sheng ZK, Hu F, Wang W, Guo Q, Chen Z, Xu X, et al. Mechanisms of Tigecycline Resistance among *Klebsiella pneumoniae* Clinical Isolates. Antimicrob Agents Chemother. 2014 Nov 1;58(11):6982.

69. Russo TA, Marr CM. Hypervirulent *Klebsiella pneumoniae*. Vol. 32, Clinical Microbiology Reviews. American Society for Microbiology; 2019.

70. Wyres KL, Lam MMC, Holt KE. Population genomics of *Klebsiella pneumoniae*. Vol. 18, Nature Reviews Microbiology. Nature Research; 2020. p. 344–59.

71. Paczosa MK, Mecsas J. *Klebsiella pneumoniae*: Going on the Offense with a Strong Defense. Microbiol Mol Biol Rev. 2016 Sep;80(3):629.

72. Patro LPP, Rathinavelan T. Targeting the Sugary Armor of *Klebsiella* Species. Front Cell Infect Microbiol. 2019 Nov 8;9:367.

73. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive web: User-Friendly capsule and lipopolysaccharide serotype prediction for *Klebsiella* genomes. J Clin Microbiol. 2018 Jun 1;56(6):197–215.

74. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of *Klebsiella* capsule synthesis loci from whole genome data. Microb Genom. 2016 Dec 1;2(12):e000102.

75. Walker KA, Treat LP, Sepúlveda VE, Miller VL. The small protein rmpd drives hypermucoviscosity in *Klebsiella pneumoniae*. mBio. 2020 Sep 1;11(5):1–14.

76. Follador R, Heinz E, Wyres KL, Ellington MJ, Kowarik M, Holt KE, et al. The diversity of *Klebsiella pneumoniae* surface polysaccharides. Microb Genom. 2016 Aug 1;2(8):e000073.

77. Chung PY. The emerging problems of *Klebsiella pneumoniae* infections: carbapenem resistance and biofilm formation. FEMS Microbiol Lett. 2016 Oct 1;363(20):219.

78. Benincasa M, Lagatolla C, Dolzani L, Milan A, Pacor S, Liut G, et al. Biofilms from *Klebsiella pneumoniae*: Matrix Polysaccharide Structure and Interactions with Antimicrobial Peptides. Microorganisms. 2016 Sep 1;4(3).

79. Cescutti P, de Benedetto G, Rizzo R. Structural determination of the polysaccharide isolated from biofilms produced by a clinical strain of *Klebsiella pneumoniae*. Carbohydr Res. 2016 Jul 22;430:29–35.

80. Bellifa S, Hassaine H, Balestrino D, Charbonnel N, M'hamedi I, Terki IK, et al. Evaluation of biofilm formation of *Klebsiella pneumoniae* isolated from medical devices at the University Hospital of Tlemcen, Algeria. Afr J Microbiol Res. 2013 Dec 3;7(49):5558–64.

81. Wang G, Zhao G, Chao X, Xie L, Wang H. The characteristic of virulence, biofilm and antibiotic resistance of *Klebsiella pneumoniae*. Vol. 17, International Journal of Environmental Research and Public Health. MDPI AG; 2020. p. 1–17.

82. Lee CR, Lee JH, Park KS, Jeon JH, Kim YB, Cha CJ, et al. Antimicrobial Resistance of Hypervirulent *Klebsiella pneumoniae*: Epidemiology, Hypervirulence-Associated Determinants, and Resistance Mechanisms. Front Cell Infect Microbiol. 2017 Nov 21;7(NOV):483.

83. Martin RM, Bachman MA. Colonization, infection, and the accessory genome of *Klebsiella pneumoniae*. Vol. 8, Frontiers in Cellular and Infection Microbiology. Frontiers Media S.A.; 2018.

84. Bachman MA, Lenio S, Schmidt L, Oyler JE, Weiser JN. Interaction of Lipocalin 2, Transferrin, and Siderophores Determines the Replicative Niche of *Klebsiella pneumoniae* during Pneumonia. mBio. 2012 Dec 31;3(6).

85. Wu H, Santoni-Rugiu E, Ralfkiaer E, Porse BT, Moser C, Høiby N, et al. Lipocalin 2 is protective against *E. coli* pneumonia. Respir Res. 2010 Jul 15;11(1):96.

86. Bachman MA, Oyler JE, Burns SH, Caza M, Lépine F, Dozois CM, et al. *Klebsiella pneumoniae* Yersiniabactin Promotes Respiratory Tract Infection through Evasion of Lipocalin 2. Infect Immun. 2011 Aug;79(8):3309.

87. Miethke M, Marahiel MA. Siderophore-Based Iron Acquisition and Pathogen Control. Microbiol Mol Biol Rev. 2007 Sep;71(3):413.

88. Fischbach MA, Lin H, Zhou L, Yu Y, Abergel RJ, Liu DR, et al. The pathogen-associated *iroA* gene cluster mediates bacterial evasion of lipocalin 2. Proc Natl Acad Sci U S A. 2006 Oct 31;103(44):16502.

89. Russo TA, Olson R, MacDonald U, Beanan J, Davidsona BA. Aerobactin, but not yersiniabactin, salmochelin, or enterobactin, enables the growth/survival of hypervirulent (hypermucoviscous) *Klebsiella pneumoniae ex vivo* and *in vivo*. Infect Immun. 2015 Aug 1;83(8):3325–33.

90. Gao Q, Wang X, Xu H, Xu Y, Ling J, Zhang D, et al. Roles of iron acquisition systems in virulence of extraintestinal pathogenic *Escherichia coli*: Salmochelin and aerobactin contribute more to virulence than heme in a chicken infection model. BMC Microbiol. 2012;12:143.

91. Catalán-Nájera JC, Garza-Ramos U, Barrios-Camacho H. Hypervirulence and hypermucoviscosity: Two different but complementary *Klebsiella spp*. phenotypes? Vol. 8, Virulence. Taylor and Francis Inc.; 2017. p. 1111–23.

92. Hsieh PF, Lin TL, Lee CZ, Tsai SF, Wang JT. Serum-Induced Iron-Acquisition Systems and TonB Contribute to Virulence in *Klebsiella pneumoniae* Causing Primary Pyogenic Liver Abscess. J Infect Dis. 2008 Jun 15;197(12):1717–27.

93. Bialek-Davenet S, Lavigne JP, Guyot K, Mayer N, Tournebize R, Brisse S, et al. Differential contribution of AcrAB and OqxAB efflux pumps to multidrug resistance and virulence in *Klebsiella pneumoniae*. Journal of Antimicrobial Chemotherapy. 2015 Jan 1;70(1):81–8.

94. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile genetic elements associated with antimicrobial resistance. Vol. 31, Clinical Microbiology Reviews. American Society for Microbiology; 2018.

95. Johnston C, Martin B, Fichant G, Polard P, Claverys JP. Bacterial transformation: distribution, shared mechanisms and divergent control. Nature Reviews Microbiology 2014 12:3. 2014 Feb 10;12(3):181–96.

96. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014 Sep 1;38(5):865.

97. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. Everyman's Guide to Bacterial Insertion Sequences. Microbiol Spectr. 2015 Apr 2;3(2).

98. Chandler M, Cruz F de la, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Microbiol. 2013 Aug;11(8):525.

99. Vandecraen J, Chandler M, Aertsen A, Houdt R van. The impact of insertion sequences on bacterial genome plasticity and adaptability. http://dx.doi.org/101080/1040841X20171303661. 2017 Nov 2;43(6):709–30.

100. Hernández-Allés S, Benedí VJ, Martínez-Martínez L, Pascual Á, Aguilar A, Tomás JM, et al. Development of Resistance during Antimicrobial Therapy Caused by Insertion Sequence Interruption of Porin Genes. Antimicrob Agents Chemother. 1999;43(4):937.

101. Naas T, Cuzon G, Truong HV, Nordmann P. Role of ISKpn7 and Deletions in *bla*$_{KPC}$ Gene Expression. Antimicrob Agents Chemother. 2012 Sep;56(9):4753.

102. Aubert D, Naas T, Héritier C, Poirel L, Nordmann P. Functional Characterization of IS1999, an IS4 Family Element Involved in Mobilization and Expression of β-Lactam Resistance Genes. J Bacteriol. 2006 Sep;188(18):6506.

103. Partridge SR, Iredell JR. Genetic Contexts of *bla*$_{NDM-1}$. Antimicrob Agents Chemother. 2012 Nov;56(11):6065.

104. Escudero JA, Loot C, Nivina A, Mazel D. The Integron: Adaptation On Demand. Microbiol Spectr. 2015 Apr 2;3(2).

105. Partridge SR, Tsafnat G. Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database. Journal of Antimicrobial Chemotherapy. 2018 Apr 1;73(4):883–90.

106. Redondo-Salvo S, Fernández-López R, Ruiz R, Vielva L, de Toro M, Rocha EPC, et al. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. Nat Commun. 2020 Dec 1;11(1):1–13.

107. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, Cruz F de la. Mobility of Plasmids. Microbiol Mol Biol Rev. 2010 Sep;74(3):434.

108. Ramsay JP, Kwong SM, Murphy RJT, Eto KY, Price KJ, Nguyen QT, et al. An updated view of plasmid conjugation and mobilization in Staphylococcus. Mob Genet Elements. 2016 Jul 3;6(4):e1208317.

109. Goessweiner-Mohr N, Arends K, Keller W, Grohmann E. Conjugation in Gram-Positive Bacteria. Microbiol Spectr. 2014 Aug 15;2(4).

110. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, 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.

111. Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M, Peto T, et al. Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. Plasmid. 2017 May 1;91:42–52.

112. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018 Aug 1;4(8).

113. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017 May 1;41(3):252–75.

114. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B, Guerra B, et al. Plasmids carrying antimicrobial resistance genes in *Enterobacteriaceae*. Journal of Antimicrobial Chemotherapy. 2018 May 1;73(5):1121–37.

115. Tatusova T, Dicuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016 Aug 19;44(14):6614–24.

116. Chen YT, Chang HY, Lai YC, Pan CC, Tsai SF, Peng HL. Sequencing and analysis of the large virulence plasmid pLVPK of *Klebsiella pneumoniae* CG43. Gene. 2004 Aug 4;337(1–2):189–98.

117. Rodrigues C, d'Humières C, Papin G, Passet V, Ruppé E, Brisse S. Community-acquired infection caused by the uncommon hypervirulent *Klebsiella pneumoniae* ST66-K2 lineage. Microb Genom. 2020;6(8):1–5.

118. 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. Nature Communications 2021 12:1. 2021 Jul 7;12(1):1–16.

119. Johnson CM, Grossman AD. Integrative and Conjugative Elements (ICEs): What They Do and How They Work. Annu Rev Genet. 2015 Nov 23;49:577.

120. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in *Klebsiella pneumoniae* populations. Microb Genom. 2018 Sep 1;4(9).

121. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. Nat Biotechnol. 2018;36(1):61–9.

122. Phelan J, de Sessions PF, Tientcheu L, Perdigao J, Machado D, Hasan R, et al. Methylation in *Mycobacterium tuberculosis* is lineage specific with associated mutations present globally. Sci Rep. 2018 Dec 1;8(1).

123. Estibariz I, Overmann A, Ailloud F, Krebes J, Josenhans C, Suerbaum S. The core genome m5C methyltransferase JHP1050 (M.Hpy99III) plays an important role in orchestrating gene expression in Helicobacter pylori. Nucleic Acids Res. 2019;47(5):2336–48.

124. Sánchez-Romero MA, Cota I, Casadesús J. DNA methylation in bacteria: From the methyl group to the methylome. Vol. 25, Current Opinion in Microbiology. Elsevier Ltd; 2015. p. 9–16.

125. Adhikari S, Curtis PD. DNA methyltransferases and epigenetic regulation in bacteria. FEMS Microbiol Rev. 2016;40(5):575–91.

126. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The Epigenomic Landscape of Prokaryotes. Fang G, editor. PLoS Genet. 2016 Feb 12;12(2):e1005854.

127. Nye TM, Jacob KM, Holleyid EK, Nevarez JM, Dawidid S, Simmons LA, et al. DNA methylation from a type I restriction modification system influences gene expression and virulence in streptococcus pyogenes. PLoS Pathog. 2019 Jun 1;15(6).

128. Wang R, Lou J, Li J. A mobile restriction modification system consisting of methylases on the IncA/C plasmid. Mob DNA. 2019 Dec 7;10(1):26.

129. Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biology 2021 22:1. 2021 Oct 18;22(1):1–33.

130. Beaulaurier J, Schadt EE, Fang G. Deciphering bacterial epigenomes using modern sequencing technologies. Nat Rev Genet. 2019 Mar 13;20(3):157–72.

131. Casselli T, Tourand Y, Scheidegger A, Arnold WK, Proulx A, Stevenson B, et al. DNA methylation by restriction modification systems affects the global transcriptome profile in *Borrelia burgdorferi*. J Bacteriol. 2018 Dec 1;200(24).

132. Pirone-Davies C, Hoffmann M, Roberts RJ, Muruvanda T, Timme RE, Strain E, et al. Genome-wide methylation patterns in *Salmonella enterica subsp. enterica* Serovars. PLoS One. 2015 Apr 10;10(4).

133. Kumar S, Karmakar BC, Nagarajan D, Mukhopadhyay AK, Morgan RD, Rao DN. N4-cytosine DNA methylation regulates transcription and pathogenesis in *Helicobacter pylori*. Nucleic Acids Res. 2018;46(7):3429–45.

134. Roberts RJ, Vincze T, Posfai JP, Macelis D. REBASE: Restriction enzymes and methyltransferases. Vol. 31, Nucleic Acids Research. 2003. p. 418–20.

135. Murray NE. Type I Restriction Systems: Sophisticated Molecular Machines (a Legacy of Bertani and Weigle). Microbiology and Molecular Biology Reviews. 2000 Jun 1;64(2):412–34.

136. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017 Mar 30;22(13):30494.

137. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. Antimicrob Agents Chemother. 2019 Nov 1;63(11).

138. 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 Dec 1;75(12):3491–500.

139. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010 Jun;7(6):461–5.

140. Hon T, Mars K, Young G, Tsai YC, Karalius JW, Landolin JM, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific Data 2020 7:1. 2020 Nov 17;7(1):1–11.

141. Sahlin K, Medvedev P. Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Nature Communications 2021 12:1. 2021 Jan 4;12(1):1–13.

142. Musich R, Cadle-Davidson L, Osier M v. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. Front Plant Sci. 2021 Apr 16;12:692.

143. Ziemann M, Kaspi A, El-Osta A. Evaluation of microRNA alignment techniques. RNA. 2016 Aug 1;22(8):1120–38.

144. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

145. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2).

146. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, Mcewen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016 Jun 20;44(11).

147. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013 Apr;10(6):563–9.

148. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017 May 1;27(5):722–36.

149. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May 1;37(5):540–6.

150. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. In: Methods in Molecular Biology. Humana Press Inc.; 2019. p. 227–45.

151. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul;30(14):2068–9.

152. Ehrlich R. Prokka database maker [Internet]. 2019 [cited 2021 May 24]. Available from: https://github.com/rehrlich/prokka_database_maker

153. Seemann T. abricate: Mass screening of contigs for antimicrobial and virulence genes [Internet]. [cited 2021 Jul 6]. Available from: https://github.com/tseemann/abricate

154. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34(Database issue).

155. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.

156. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan 1;32(1):268–74.

157. Katoh K. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002 Jul 15;30(14):3059–66.

158. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.

159. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 2005;15(2):330–40.

160. Gregory TR. Understanding Evolutionary Trees. Evolution: Education and Outreach. 2008 Feb 26;1(2):121–37.

161. Philippe H, Brinkmann H, Lavrov D v., Littlewood DTJ, Manuel M, Wörheide G, et al. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol. 2011 Mar;9(3).

162. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121–3.

163. Stott CM, Bobay LM. Impact of homologous recombination on core genome phylogenies. BMC Genomics. 2020 Dec 1;21(1):1–10.

164. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLoS Comput Biol. 2015;11(2).

165. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015 Feb 18;43(3):e15.

166. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data. J Bacteriol. 2004 Mar;186(5):1518.

167. Tan M, Long H, Liao B, Cao Z, Yuan D, Tian G, et al. QS-Net: Reconstructing phylogenetic networks based on quartet and sextet. Front Genet. 2019;10(JUN):607.

168. Dagan T. Phylogenomic networks. Trends Microbiol. 2011 Oct 1;19(10):483–91.

169. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. mBio. 2014 Nov 3;5(6).

170. Wickelmaier F. An Introduction to MDS. 2003;

171. Lin T, Zha H. Riemannian manifold learning. IEEE Trans Pattern Anal Mach Intell. 2008 May;30(5):796–809.

172. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv [Internet]. 2018 Feb 9 [cited 2021 Apr 22]; Available from: http://arxiv.org/abs/1802.03426

173. van der Maaten L, Hinton G. Visualizing Data using t-SNE. Vol. 9, Journal of Machine Learning Research. 2008.

174. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–30.

175. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet. 2021 Jan 1;66(1):85–91.

176. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 Jan 1;37(1):38–47.

177. Yang Y, Sun H, Zhang Y, Zhang T, Gong J, Wei Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep. 2021 Jul 27;36(4).

178. Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013;7819 LNAI(PART 2):160–72.

179. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nature Reviews Genetics 2016 18:1. 2016 Nov 14;18(1):41–50.

180. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018 Feb 1;14(2).

181. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics. 2018 Dec 15;34(24):4310–2.

182. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. 9th Python in Science Conference [Internet]. 2010 [cited 2022 Mar 1]; Available from: http://statsmodels.sourceforge.net/

183. Perdigão J, Modesto A, Pereira AL, Neto O, Matos V, Godinho A, et al. Whole-genome sequencing resolves a polyclonal outbreak by extended-spectrum beta-lactam and carbapenem-resistant *Klebsiella pneumoniae* in a Portuguese tertiary-care hospital. Microb Genom. 2020 Apr 1;mgen000349.

184. Parish T, Stoker NG, van Soolingen D, de Haas PEW, Kremer K. Restriction Fragment Length Polymorphism Typing of Mycobacteria. In: *Mycobacterium Tuberculosis* Protocols. Humana Press; 2003. p. 165–203.

185. Clark TA, Lu X, Luong K, Dai Q, Boitano M, Turner SW, et al. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biol. 2013;11(1):4.

186. Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44(D1):D7–19.

187. National Library of Medicine. The NCBI Pathogen Detection Project. 2016.

188. Loraine J, Heinz E, de Sousa Almeida J, Milevskyy O, Voravuthikunchai SP, Srimanote P, et al. Complement Susceptibility in Relation to Genome Sequence of Recent *Klebsiella pneumoniae* Isolates from Thai Hospitals . mSphere. 2018 Nov 7;3(6).

189. Russo TA, MacDonald U, Hassan S, Camanzo E, LeBreton F, Corey B, et al. An Assessment of Siderophore Production, Mucoviscosity, and Mouse Infection Models for Defining the Virulence Spectrum of Hypervirulent *Klebsiella pneumoniae*. mSphere. 2021 Mar 24;6(2).

190. Struve C, Roe CC, Stegger M, Stahlhut SG, Hansen DS, Engelthaler DM, et al. Mapping the evolution of hypervirulent *Klebsiella pneumoniae*. mBio. 2015 Jul 21;6(4).

191. Liu C, Dong N, Chan EWC, Chen S, Zhang R. Molecular epidemiology of carbapenem-resistant *Klebsiella pneumoniae* in China, 2016-20. Lancet Infect Dis. 2022 Feb 1;22(2):167–8.

192. Liao W, Liu Y, Zhang W. Virulence evolution, molecular mechanisms of resistance and prevalence of ST11 carbapenem-resistant *Klebsiella pneumoniae* in China: A review over the last 10 years. J Glob Antimicrob Resist. 2020 Dec 1;23:174–80.

193. Wyres KL, Nguyen TNT, Lam MMC, Judd LM, van Vinh Chau N, Dance DAB, et al. Genomic surveillance for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from South and Southeast Asia. Genome Med. 2020 Jan 16;12(1).

194. Chen L, Mathema B, Pitout JDD, DeLeo FR, Kreiswirth BN. Epidemic *Klebsiella pneumoniae* ST258 is a hybrid strain. mBio. 2014 Jun 24;5(3).

195. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, 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 U S A. 2015 Jul 7;112(27):E3574–81.

196. Gorrie CL, da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. Lancet Microbe. 2021 Nov 1;2(11):e575–83.

Appendix 1. Supplementary materials: Methylation analysis of *Klebsiella pneumoniae* from Portuguese hospitals.

Figure S1, The motifs analysed. Comparison between the generality of all methylation motifs identified by SMRT Analysis and the rate of motif methylation. Only motifs with >60% share of methylation were selected for further analysis. The motifs that cover >100% of chromosome were very general non-palindromic motifs. The total length of some motifs exceeds the chromosome length due to occurrence on both DNA strands. The three analysed motifs with a share of methylation below 50% are GATC and both partners of GCAYN$_5$GTT (from Kp2564). The low methylation rate is due to low sequencing coverage.

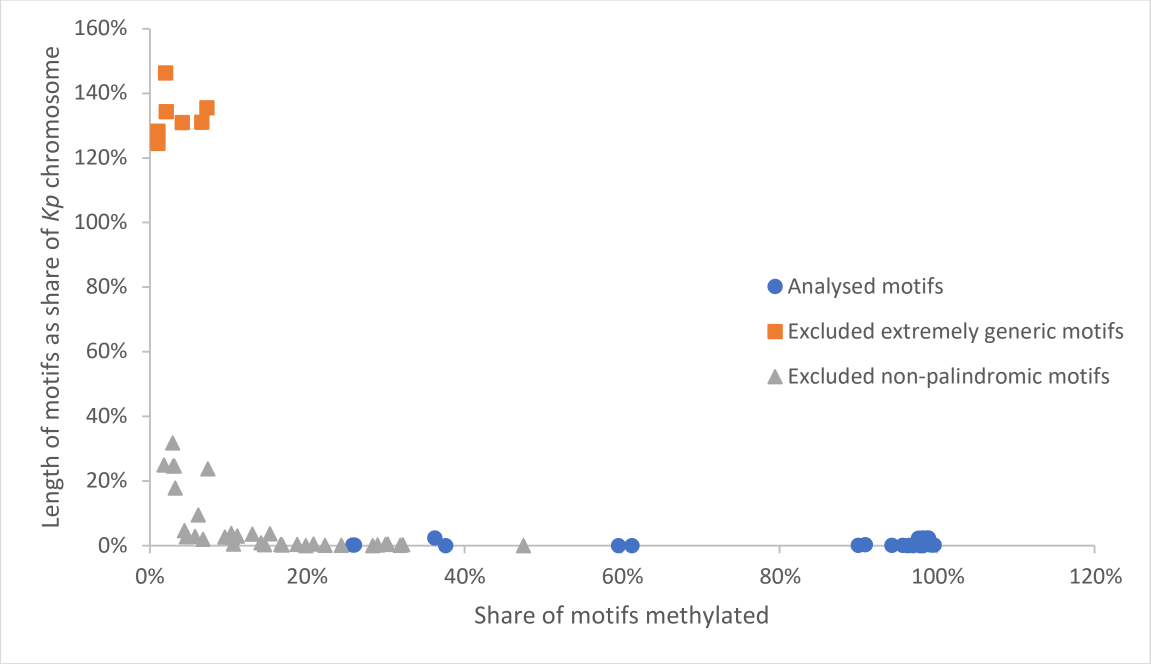


Figure S2, GATC motifs inter-pulse duration (IPD) ratio for the *fosA* centred chromosomal regions (see **Figure 3**). The four outliers, representing unmethylated GATC downstream of *fosA* are visible at the bottom of the whiskers for Kp1675, Kp2209, Kp2958 and Kp3860. The single outlier of Kp1208 is less visible. The two samples without outliers Kp1363 and Kp1264 do not have the GATC motif downstream of *fosA* due to an insertion of endonuclease (see **Figure 3**).

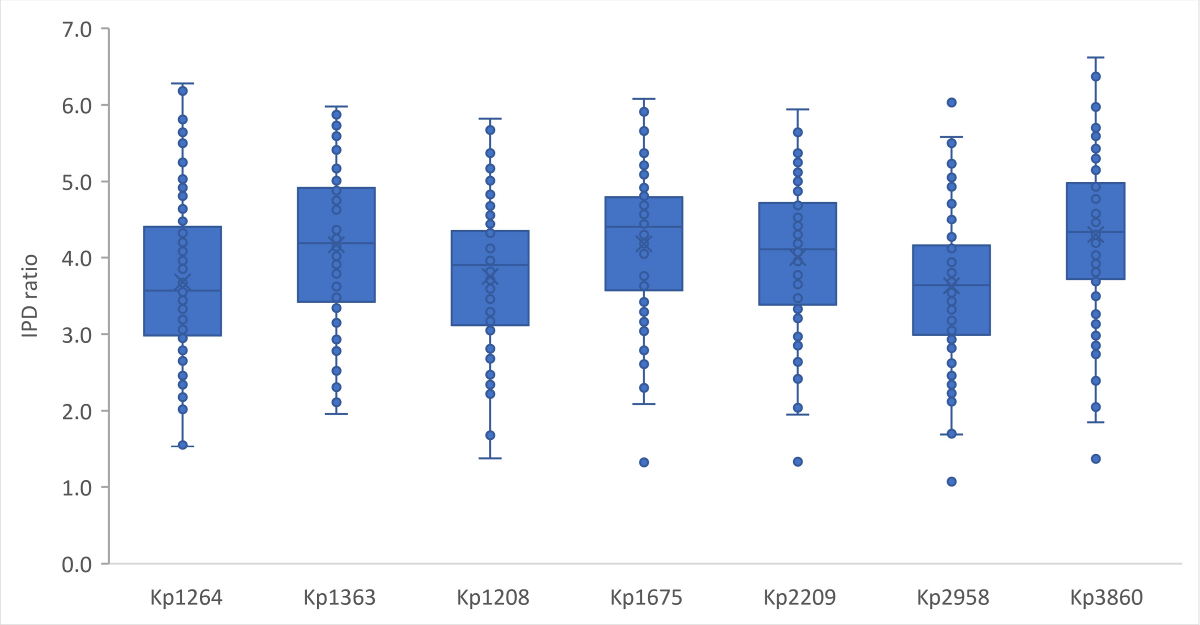
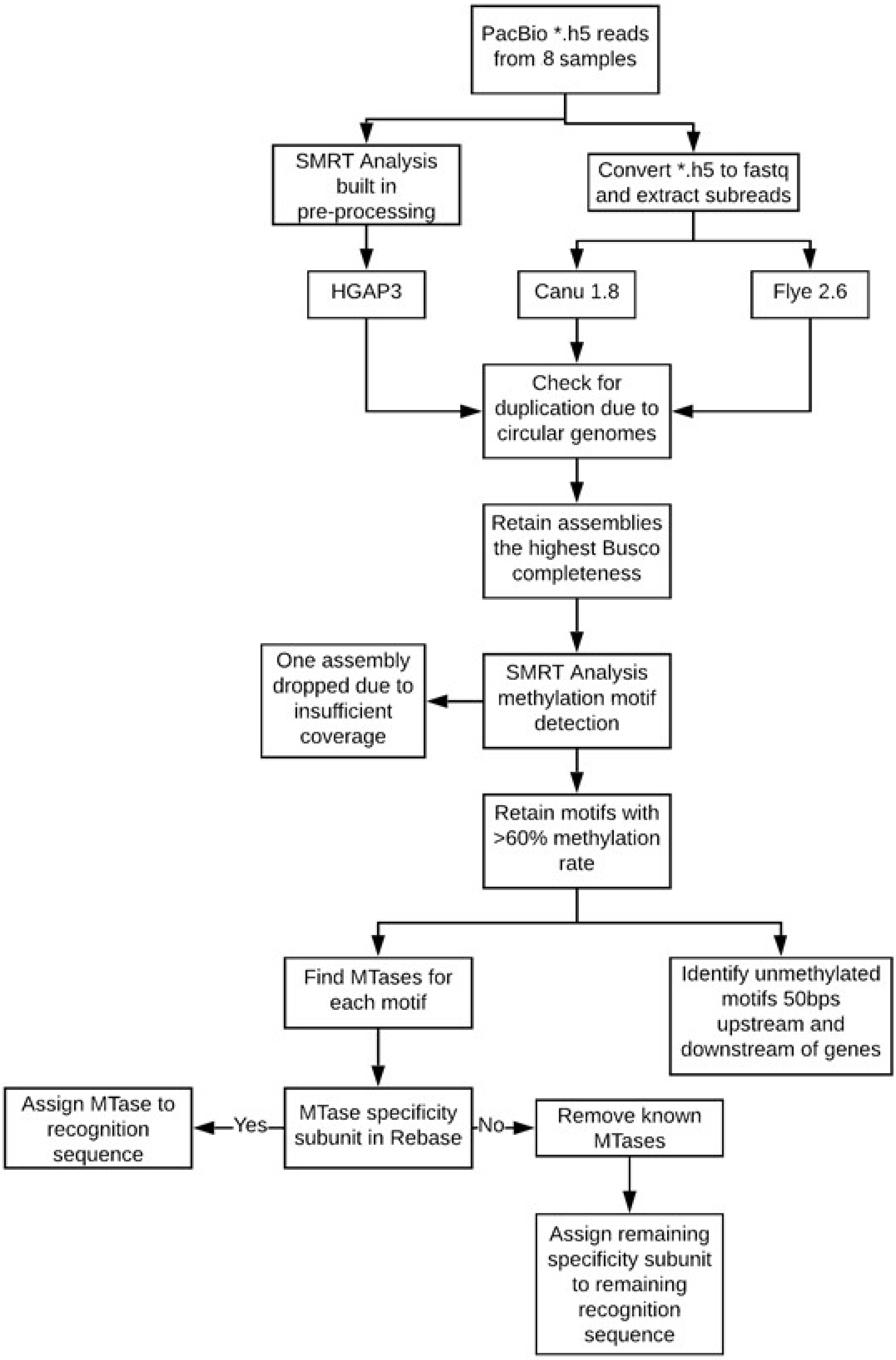


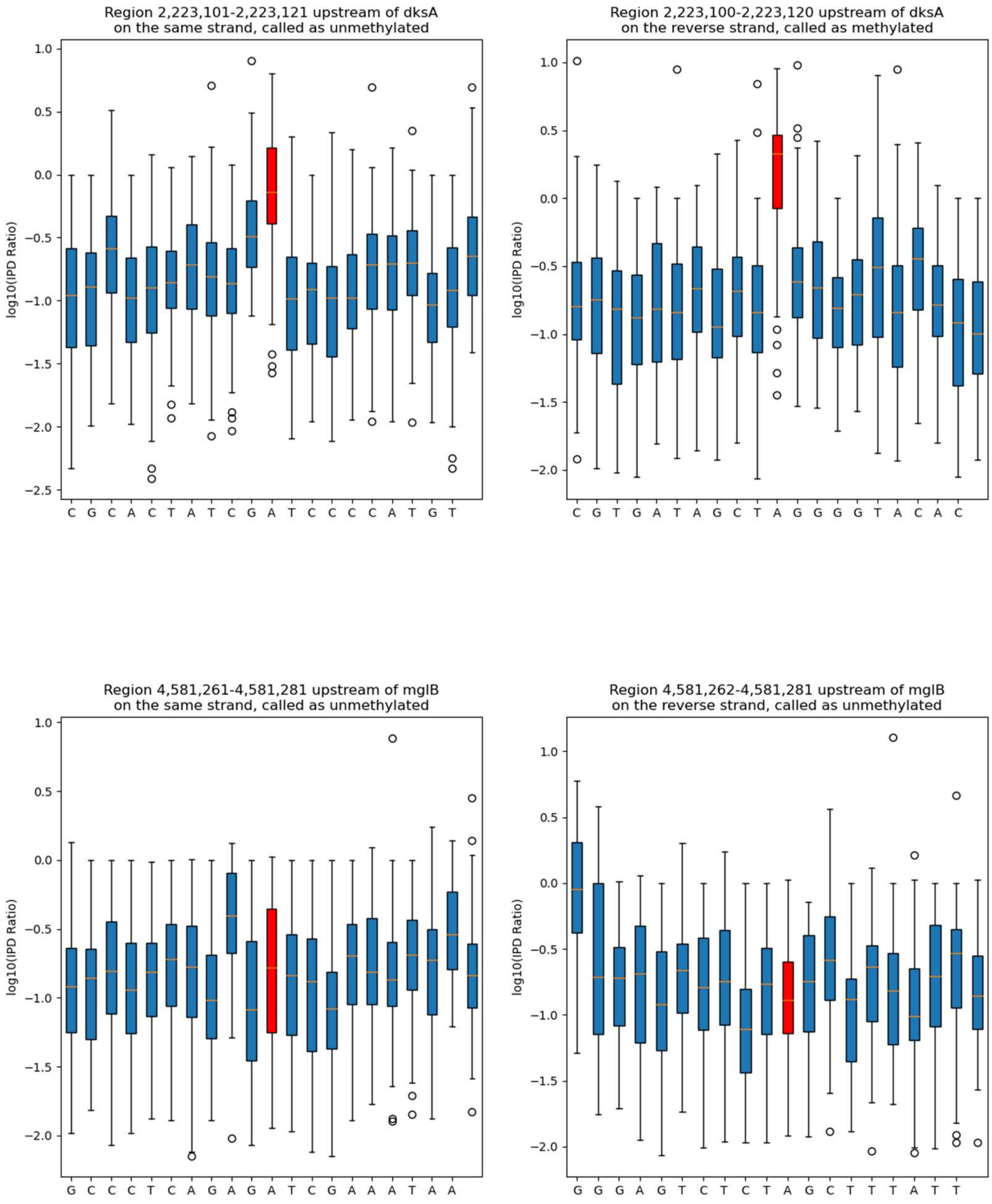
Figure S3, The analytical pipeline



PacBio *.h5 reads from 8 samples
SMRT Analysis built in pre-processing
Convert *.h5 to fastq and extract subreads
HGAP3
Canu 1.8
Flye 2.6
Check for duplication due to circular genomes
Retain assemblies the highest Busco completeness
One assembly dropped due to insufficient coverage
SMRT Analysis methylation motif detection
Retain motifs with >60% methylation rate
Find MTases for each motif
Identify unmethylated motifs 50bps upstream and downstream of genes
Assign MTase to recognition sequence
Yes
MTase specificity subunit in Rebase
No
Remove known MTases
Assign remaining specificity subunit to remaining recognition sequence

Figure S4, Log$_{10}$(IPD ratios) around GATC motifs upstream of (a) *dksA* (BAH61791.1) and (b) *mglB* (VEC00318.1) genes in the Kp3860 isolate. Red is adenine in GATC motif. The difference between methylated and unmethylated calls in *dksA* **(a)** is not large and is driven by different means. Some outliers are also visible, and some are hidden by long whiskers. By contrast, in *mglB* **(b)** adenine does not stand out from other bases.

(a)



(b)

Table S1, Antimicrobial Resistance genotype and phenotype

ID	NCBI accession	CTX-M-15	OXA-1	KPC-3	AMC	FOX	CTX	CAZ	IPM	GM	CIP	FOS
Kp1363	CAESWS000000000.1				S	S	S	S	S	S	S	S
Kp1208	CAESWW000000000.1				R	S	S	R	S	R	S	S
Kp1264	CAESWT000000000.1	Tn3 IncFIA(HI1)	Tn3 IncFIA(HI1)		R	R	R	R	S	R	R	R
Kp1675	CAESWY000000000.1	Tn3	Tn3		R	S	R	R	S	R	S	S
Kp2209	CAESWV000000000.1	IncFIB(K)			R	S	R	R	S	S	S	S
Kp2564	CAESWU000000000.1			IncFIA	R	R	R	R	R	R	R	S
Kp2958	CAESWX000000000.1				R	R	R	R	I	R	R	S
Kp3860	CAESWZ000000000.1	IncFIB(K)	IncFIB(K)		R	S	R	R	S	R	R	S

S = sensitive, R = resistant, AMC= amoxicillin + clavulanic acid, FOX = cefoxitin, CTX = cefotaxime, CAZ = ceftazidime, IPM = imipenem, GM = gentamicin, CIP = ciprofloxacin, FOS = fosfomycin

Table S2. Comparison of abundance ratio (AR) of recognition motifs on chromosomes (contigs >1Mbp) and putative extra-chromosomal genetic elements (PEGEs; contigs <1Mbp).

Group 1	**Group 2**	**Mean AR % Group 1**	**Mean AR % Group 2**	**Wilcoxon rank sum test P-value**
Chromosome type I recognition motifs native to assembly	Chromosome type I recognition motifs from other assemblies	0.192	0.234	0.515
PEGE type I recognition motifs native to assembly	PEGE type I recognition motifs from other assemblies	0.183	0.268	0.187
Chromosome type I recognition motifs	PEGE elements (>10kbp) type I recognition motifs	0.230	0.262	0.234
GATC motifs on chromosome	GATC motifs on PEGEs	2.25	1.45	0.000003

Table S3. Of the 3,584 *K. pneumoniae* genomes analysed, only these 23 did not have gene *syrM1* immediately upstream of *fosA* gene.

Assembly ID	Sequence type	Country
GCA_000529745.1	ST3695	Austria
GCA_000529945.1	ST15-3LV	Austria
GCA_000529425.1	ST3827-3LV	Austria
GCA_003095515.1	ST258	Brazil
GCA_004127575.1	ST256	China
GCA_002173825.1	ST25	China
GCA_900516965.1	ST15	Hungary
GCA_004145895.1	ST231	India
GCA_900181455.1	ST48	Pakistan
GCA_001316495.2	ST101	South Africa
GCA_001316565.3	ST101	South Africa
GCA_001316645.2	ST101	South Africa
GCA_001316785.2	ST101	South Africa:
GCA_001316895.2	ST101	South Africa
GCA_001316985.2	ST101	South Africa
GCA_002510005.1	ST3800	South Africa
GCA_002510215.1	ST1552	South Africa
GCA_002522955.1	ST101	South Africa
GCA_002856415.1	ST199	USA
GCA_002856565.1	ST133	USA
GCA_002856885.1	ST968-4LV	USA
GCA_002857345.1	ST298	USA
GCA_001066585.1	ST3602	USA

Data S1, full list of genes with identified upstream or downdstream unmethylated motifs

			# of unmethylated instances in a sample							# unmethylated samples (might be missing motif)
Location	product	RefSeq	WBB1001 1675	WBB1005 Kp1264	WBB1003 Kp1208	WBB1000 Kp1363	WBB1004 Kp2958	WBB998 Kp2209	WBB997 Kp3860	
up	D-ribose transporter subunit RbsB	VEC00318.1	1	1	1	1	1	1	1	7
up	Branched-chain amino acid transport protein azlC	AHM82117.1	1	1	1	0	1	1	1	6
up	D-arabinitol dehydrogenase	BAH64332.1	0	1	1	1	1	1	1	6
up	DnaK suppressor protein	BAH61791.1	1	0	1	1	1	1	1	6
up	BCCT family transporter	AHM80284.1	0	1	1	1	1	1	0	5
down	FosA family fosfomycin resistance glutathione transferase	QBH08895.1	1	0	1	0	1	1	1	5
down	DUF1145 family protein	AHM77076.1	1	0	1	1	0	1	0	4
down	IS3 family transposase	VEC38624.1	0	0	1	0	1	1	1	4
up	Transcriptional regulators of sugar metabolism	VEC00015.1	0	1	1	0	1	0	1	4
up	transcriptional antiterminator BglG	BAH64373.1	0	1	1	0	1	0	0	3
up	LysR family transcriptional regulator	VED54907.1	1	0	1	0	1	0	0	3
up	outer membrane receptor FepA	VEB98368.1	0	1	1	0	1	0	0	3
up	pectin degradation protein kdgF	VEC02571.1	0	1	0	1	1	0	0	3
up	putative DeoR-type transcriptional regulator	BAH64418.1	0	1	1	0	0	1	0	3
N/A	putative protein		1	0	3	0	1	1	0	4
up	Competence/damage-inducible protein CinA	VEC28104.1/ VEC06933.1	1	0	0	0	0	0	1	2
down	Dienelactone hydrolase family protein	AHM79952.1	2	0	1	0	0	0	0	2
up	efflux RND transporter periplasmic adaptor subunit	AWD06510.1	0	0	1	0	1	0	0	2
down	Fructose repressor FruR	SQI85527.1	1	0	0	0	0	0	1	2

down	metal ABC transporter substrate-binding protein	AGT22970.1	0	1	0	0	0	1	0	2
up	inner membrane protein YdgG	ACI11713.1	1	0	1	0	0	0	0	2
up	L-fucose isomerase	VEC02477.1	0	1	0	0	1	0	0	2
down	L-fucose:H+ symporter permease	ATU18929.1	0	1	0	0	1	0	0	2
both	protein YobH	VEC00531.1	0	0	2	0	0	1	0	2
down	putative DMT superfamily transporter inner membrane protein	AEG99519.1	1	0	0	0	0	1	0	2
down	relaxase	AUD32897.1	0	0	0	0	0	1	1	2
up	3-(3-hydroxy-phenyl)propionate hydroxylase	AHM79080.1	0	0	1	0	0	0	0	1
both	3-oxoacyl-ACP synthase	VEC55131.1	2	0	0	0	0	0	0	1
up	5-methyltetrahydropteroyltriglutamate/ homocysteine S-methyltransferase	VEB03045.1	1	0	0	0	0	0	0	1
down	adenosine-3'(2')%2C5' -bisphosphate nucleotidase	AEG96821.1	0	0	1	0	0	0	0	1
down	Alkaline phosphatase isozyme conversion protein precursor	VEC02421.1	0	0	0	1	0	0	0	1
up	alpha-galactosidase	QAZ97860.1	0	1	0	0	0	0	0	1
up	aspartate kinase III	AEG96592.1	0	0	0	1	0	0	0	1
up	ATP-binding protein	QBH46725.1	2	0	0	0	0	0	0	1
up	baseplate assembly protein	AIJ39130.1	1	0	0	0	0	0	0	1
down	bifunctional nitric oxide dioxygenase/dihydropteridine reductase 2	AEG95105.1	0	1	0	0	0	0	0	1
down	bile acid/Na+ symporter family transporter	AGT23254.1	0	0	0	1	0	0	0	1
down	BolA protein	AHM81027.1	1	0	0	0	0	0	0	1
up	chromosome partitioning protein ParB	AZI04767.1	0	1	0	0	0	0	0	1

up	cyclopropane-fatty-acyl-phospholipid synthase	BAS41165.1	0	1	0	0	0	0	0	1
up	D-arabinitol 4-dehydrogenase	VEC01152.1	1	0	0	0	0	0	0	1
down	DNA-invertase	VEC99321.1	0	0	0	0	0	1	0	1
up	Exopolysaccharide biosynthesis protein	VEC12215.1	1	0	0	0	0	0	0	1
down	Gluconolactonase	AHM78164.1	0	0	1	0	0	0	0	1
down	Glyoxalase-like domain protein	VEB98524.1	0	0	0	0	0	0	1	1
down	multidrug ABC transporter permease/ATP-binding protein	AOA94660.1	0	0	0	1	0	0	0	1
up	ubiquinone-dependent pyruvate dehydrogenase		1	0	0	0	0	0	0	1
up	hypothetical protein		2	0	0	0	0	0	0	1
down	DUF4158 domain-containing protein		1	0	0	0	0	0	0	1
down	tRNA (N6-threonylcarbamoyladenosine(37)-N6)-methyltransferase TrmO		1	0	0	0	0	0	0	1
up	Prophage P2 OGR protein	CDO13729.1	3	0	0	0	0	0	0	1
up	putative transcriptional regulator		0	0	1	0	0	0	0	1
down	biofilm regulator BssR	ABR76304.1	0	0	0	0	1	0	0	1
down	DNA damage-inducible protein D	CDO11786.1	0	0	0	0	0	0	1	1
down	MFS transporter (Methylynomycin resistance)	BAH65098.1	0	0	0	0	0	1	0	1
up	immunity repressor	VED52142.1	0	0	0	0	0	0	1	1
down	inorganic diphosphatase	VEB99626.1	1	0	0	0	0	0	0	1
up	integrase	AEW91928.1	1	0	0	0	0	0	0	1
down	integrating conjugative element protein PilL%2C PFGI-1 class	VED56414.1	0	0	0	1	0	0	0	1
down	IS21-like element ISEc12 family transposase	QBH43493.1	2	0	0	0	0	0	0	1
down	LysR family transcriptional regulator YneJ	VEC78603.1	0	0	1	0	0	0	0	1

up	Macrophage infectivity potentiator-related protein	VED45250.1	0	0	1	0	0	0	0	1
down	Maltodextrin phosphorylase	AHM77127.1	1	0	0	0	0	0	0	1
down	metal chaperone	VEB00693.1	0	0	0	1	0	0	0	1
down	Methyl viologen resistance protein smvA	VEB99700.1	1	0	0	0	0	0	0	1
down	NLP/P60 protein	AEG98321.1	0	0	0	0	1	0	0	1
down	outer membrane porin HofQ	AGT22205.1	1	0	0	0	0	0	0	1
up	para-aminobenzoate synthase component I	BAH64055.1	0	0	1	0	0	0	0	1
up	peptidase M28	VEB83718.1	0	0	0	1	0	0	0	1
down	phage portal family protein	AWF47745.1	1	0	0	0	0	0	0	1
up	Phosphonate ABC transporter ATP-binding protein (TC 3.A.1.9.1)	VED43404.1	0	0	1	0	0	0	0	1
up	product=hypothetical protein	ATO04573.1	0	0	0	0	1	0	0	1
down	Protein YjgK%2C linked to biofilm formation	VEC04079.1	1	0	0	0	0	0	0	1
up	putative family 32 glycoside hydrolase	BAH65040.1	0	0	0	0	1	0	0	1
down	putative Nudix hydrolase NudL	protein motif: HAMAP: MF_01592	0	0	1	0	0	0	0	1
up	response regulator in two-component regulatory system with PhoQ	BAH62844.1	0	0	0	0	0	0	1	1
down	Rhodanese-related sulfurtransferase	AHM80228.1	0	0	0	1	0	0	0	1
up	Rop family plasmid primer RNA-binding protein	ATR77797.1	0	1	0	0	0	0	0	1
down	secretion protein HlyD	AEG96162.1	0	0	0	1	0	0	0	1
up	tail protein	AIJ39139.1	1	0	0	0	0	0	0	1
down	tonB-dependent receptor yncD	VTO27455.1	0	0	0	1	0	0	0	1
down	transcriptional activator for 3-hydroxyphenylpropionate degradation	BAH63840.1	0	0	1	0	0	0	0	1

up	transcriptional regulator	VEC03025.1	1	0	0	0	0	0	0	1
up	Transcriptional regulator%2C TetR family	VEC00153.1	0	0	0	0	1	0	0	1
down	Transposase	VEC38481.1	0	0	1	0	0	0	0	1
down	tRNA-Gly(gcc)		0	0	0	1	0	0	0	1
down	YoaH family protein	ATQ98873.1	0	0	1	0	0	0	0	1

Appendix 2. Supplementary materials: Genomic epidemiological analysis of *Klebsiella pneumoniae* from Portuguese hospitals reveals insights into circulating antimicrobial resistance.

Figure S1. Location of Kp genes used for phylogenetic reconstructions in **Figure 1**, with their presence across sequence types (STs).

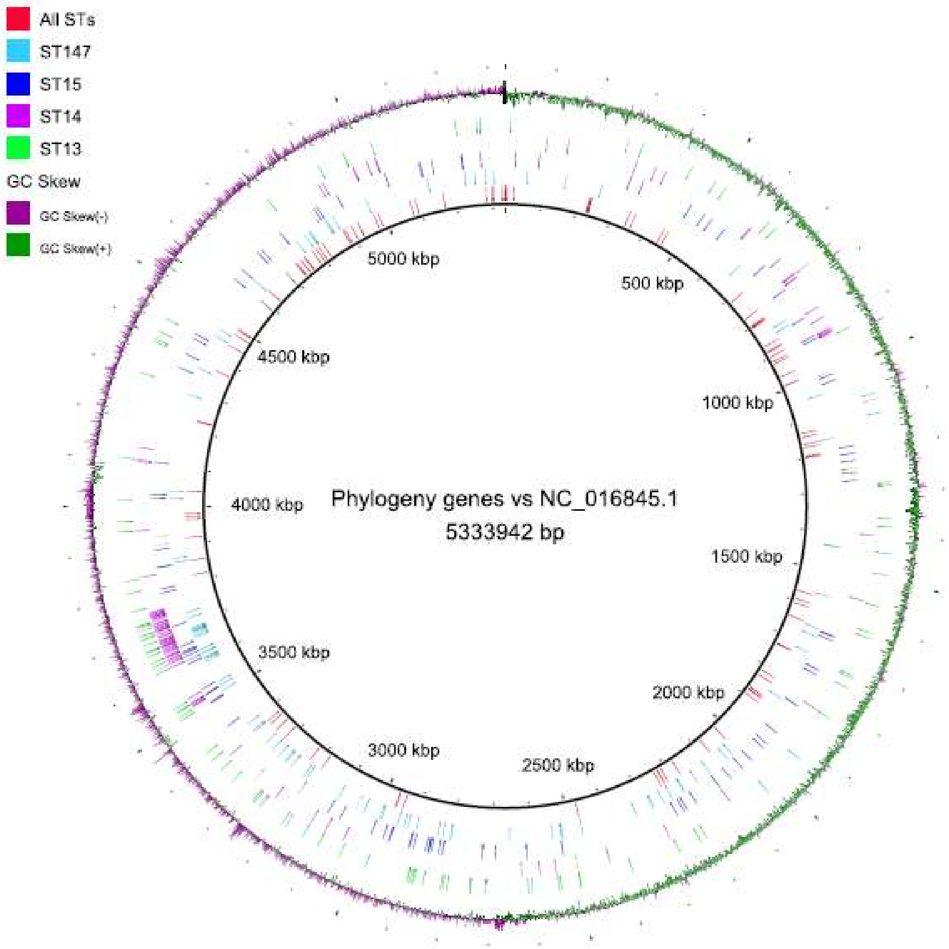


Figure S2. Multi-locus sequence typing (MLST) allele distance between unique sequence types.

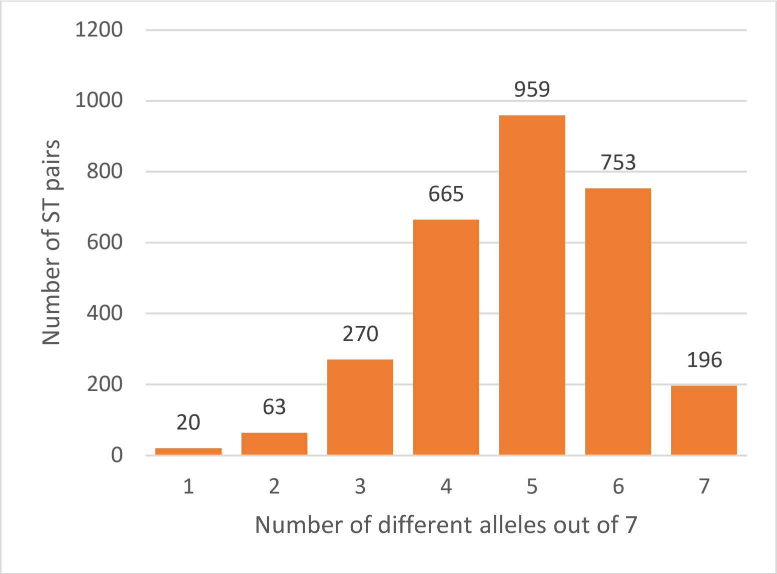


Figure S3. Maximum likelihood phylogenetic tree for all isolates (n=509) using core genes; O-loci are inferred O serotypes; ST sequence types; RC replicon cluster; boot = bootstrap support %; AMR antimicrobial resistance.

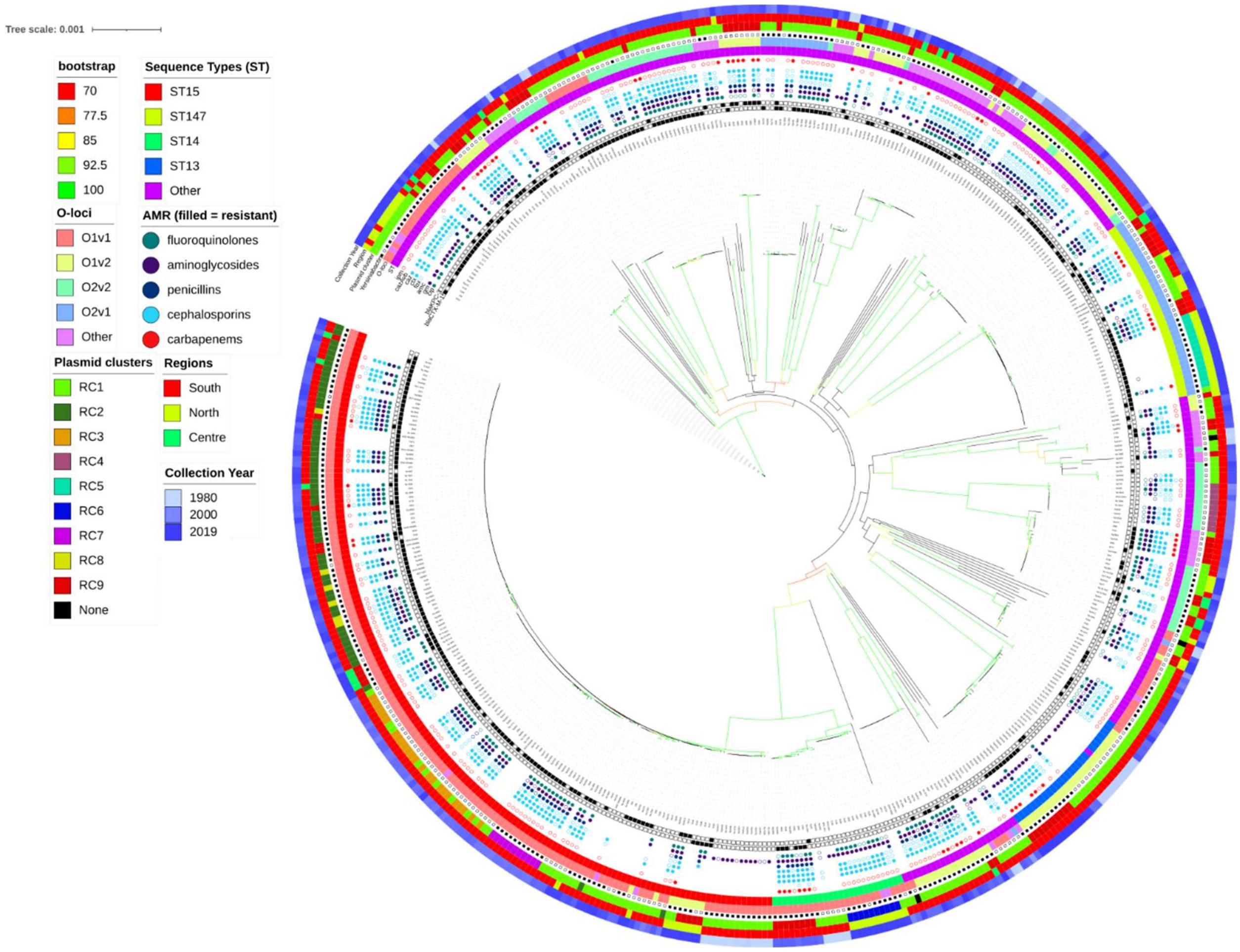
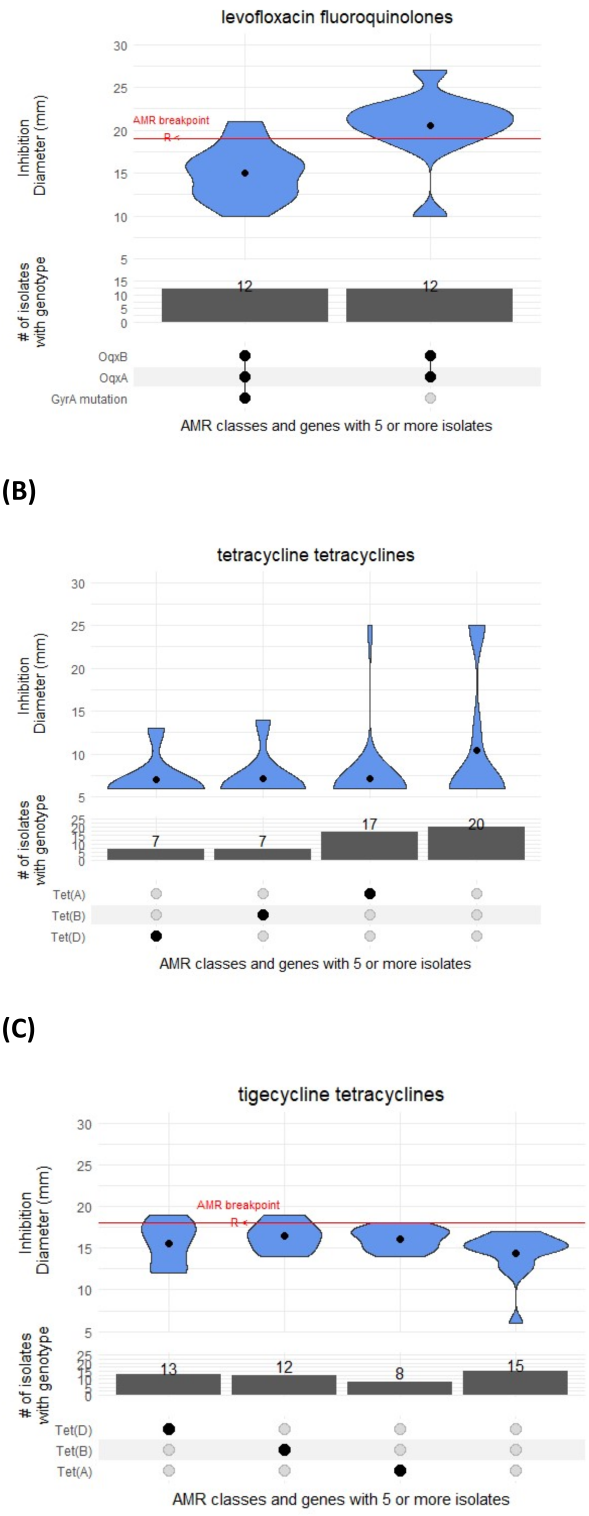


Figure S4. Inhibition zone diameters for selected antimicrobials **(A**) Levofloxacin; **(B)** Tetracycline; **(C)** Tigecycline. Note, EUCAST does not have a Kp breakpoint for tetracycline. For tigecycline, EUCAST v11.0 does not define a Kp breakpoint, so *E. coli* breakpoint is used instead.

(A)

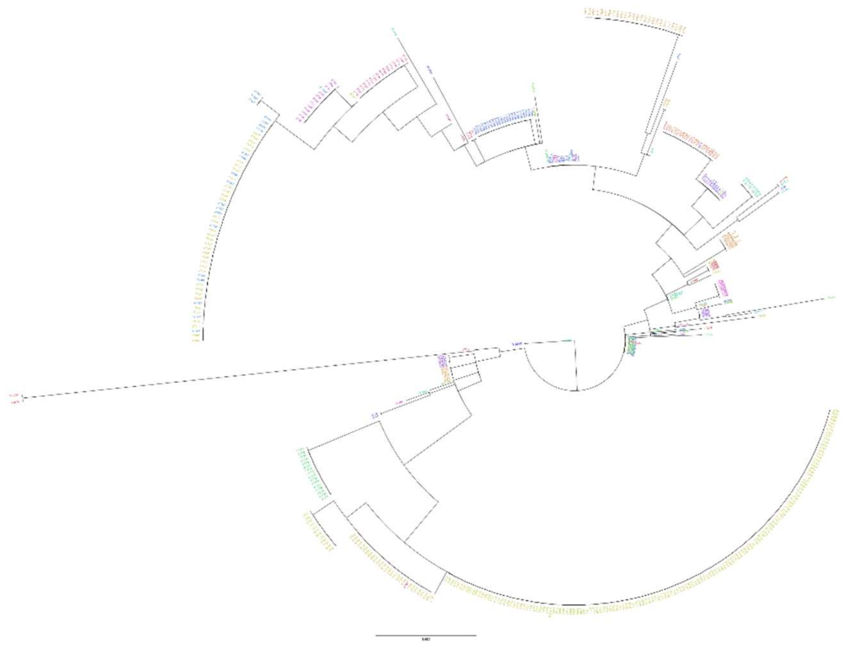


(B)

(C)

Figure S5. Maximum likelihood phylogenetic tree for all isolates (n=509) using the: **(A)** *parC* gene; (**B)** *gyrA* gene.

(A)



(B)

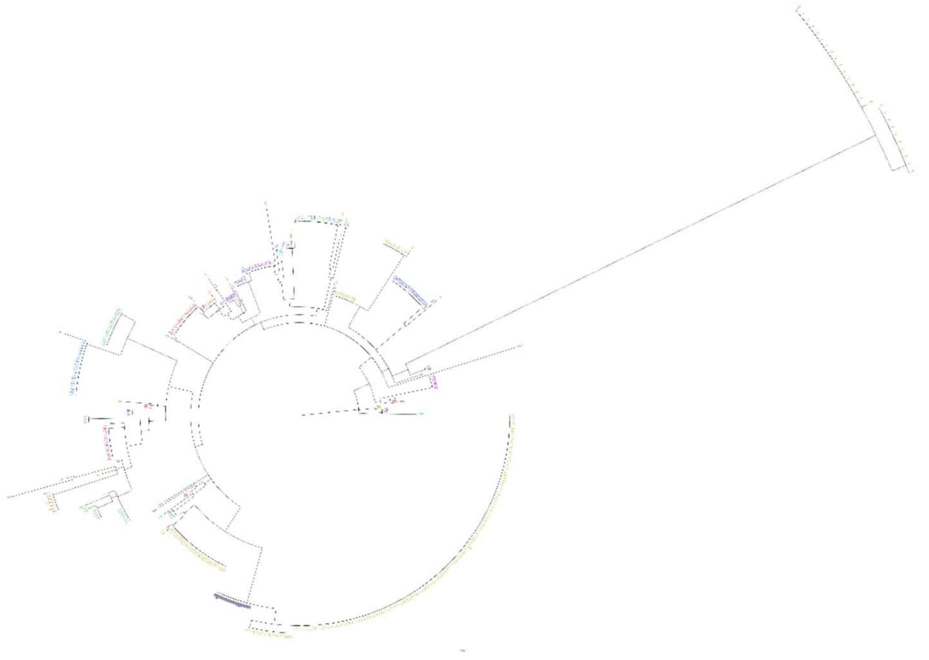


Table S1. Resistance phenotypes of isolates undergoing drug susceptibility tests (DSTs)

Antimicrobial Family	Name	Abbrev.	DST N	DST Resistance N	DST Resistance %
Aminoglycosides	gentamicin	gm	356	313	88
	amikacin	ank	29	16	55
Carbapenems	imipenem	ipm	296	68	23
	meropenem	mem	45	32	71
	doripenem	dor	12	12	100
Cephalosporins 2nd	cefoxitin	fox	366	157	43
Cephalosporins 2nd	cefuroxime	cxm	21	14	67
Cephalosporins 3rd	cefotaxime	ctx	362	279	77
Cephalosporins 3rd	ceftazidime	caz	357	303	85
Cephalosporins 3rd	ceftazidime-avibactam	cazAvb	45	1	2
Cephalosporins 3rd	ceftriaxone	cro	28	22	79
Cephalosporins 4th	cefepime	fep	39	35	90
Fluoroquinolones	ciprofloxacin	cip	344	258	75
	levofloxacin	lev	35	20	57
	norfloxacin	nor	23	13	57
Miscellaneous agents	fosfomycin	fos	27	18	67
monobactam	aztreonam	atm	86	77	90
Penicillins	amoxicillin-clavulanic acid	amc	311	292	94
	amoxicillin	amx	56	56	100
	piperacillin	pip	45	45	100
	ticarcillin-clavulanic acid	tcc	34	34	100
	ticarcillin	tic	34	34	100
Tetracyclines	tigecycline	tig	48	40	83

Data S1, Count of plasmid replicon hashes across replicon clusters.

Data S2, Metadata for all isolates in the study including isolate origin and collection date, AMR phenotype and Kleborate genotyping results.

Appendix 3. Supplementary materials: Genomic analysis of hypervirulent *Klebsiella pneumoniae* reveals potential genetic markers for differentiation from classical strains.

Figure S1. Clustering of isolates based on accessory genome demonstrates that accessory genes are linked to sequence type (ST) and not geography with ST23 being a tight cluster. Each point is an isolate and axis are dimensionless.

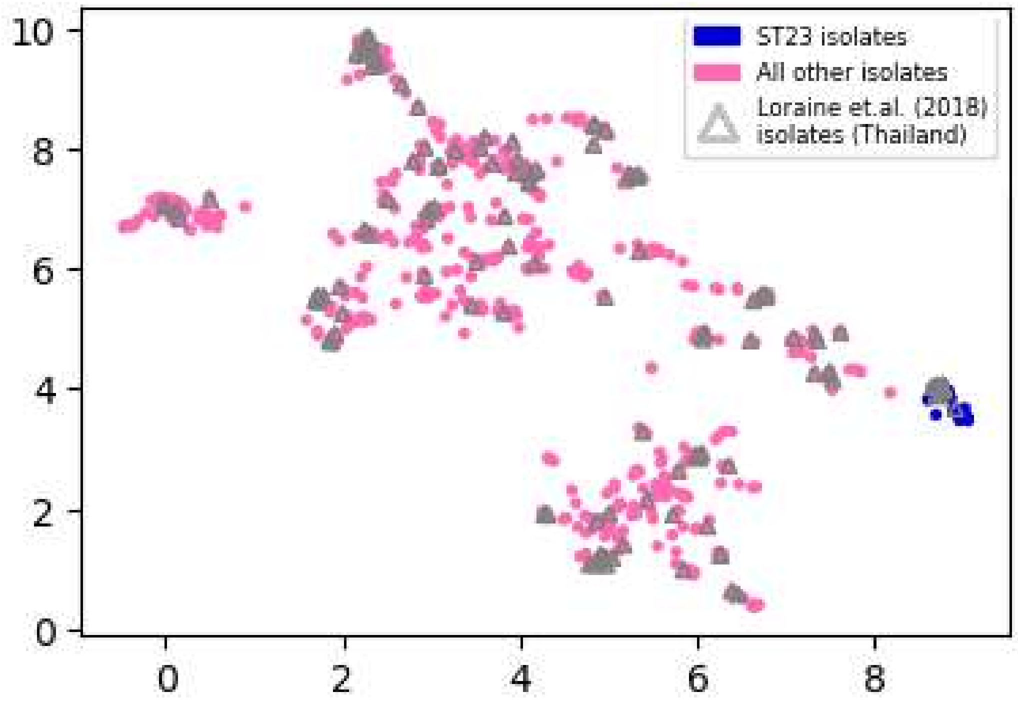


Figure S2. **(A)** This figure is the same as **Figure S1A**, except genes are coloured by clusters identified by DBSCAN algorithm. **(B)** The location of clusters of genes in **(A)** on a *K. pneumoniae* virulence plasmid pLVPK. Genes from same clusters occur together on the plasmid, but sometimes they are split-up by genes from another cluster.

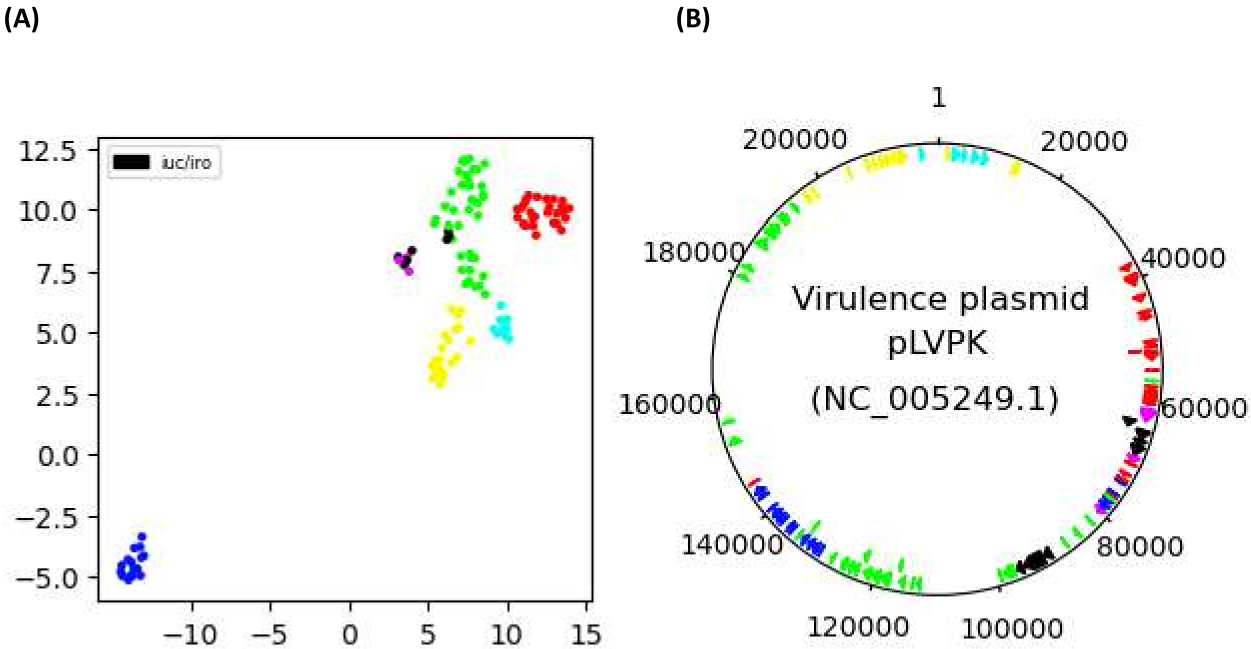
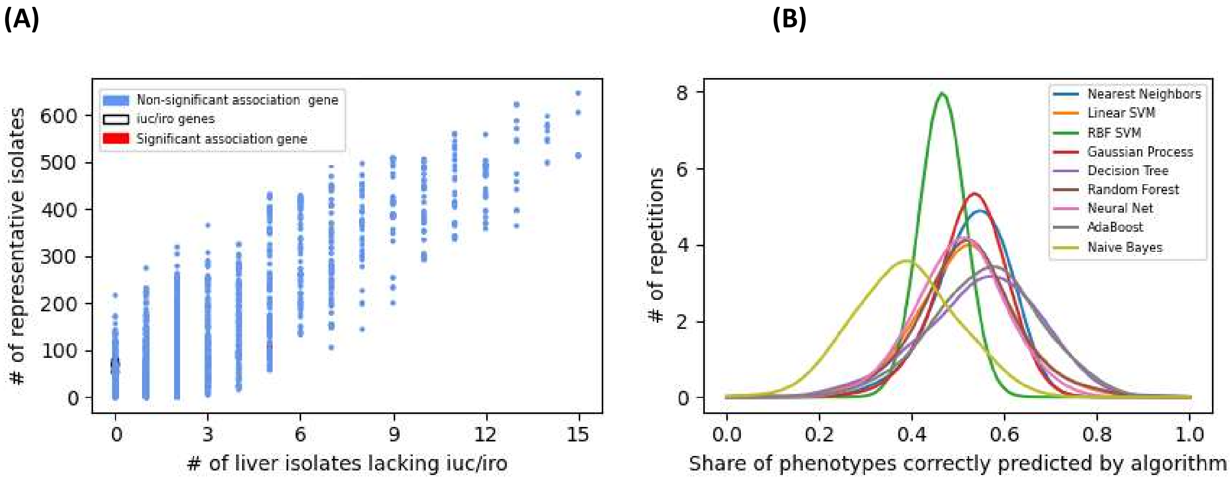


Figure S3. **(A)** Frequency of accessory genome genes in liver samples lacking *iro* and *iuc* loci versus representative dataset. **(B)** Phenotype prediction accuracy of different machine learning algorithms based on 100 iterations per algorithm. Each iteration was based on all 15 liver isolates lacking *iro* and *iuc* as well as 15 isolates randomly chosen from representative dataset.



Data S1. Summary of isolates used in the study (Text file)

Data S2. Fasta sequences for all genes in **Figure 3B** and **Data S3**

Data S3. List of all genes in **Figure 3B**

ID	X1	X2	Description	Count Liver ExST23	Count Liver ST23	Count NonLiver	-LOG10(pvalue)
B385338	6.34	8.95	hypothetical protein	34	24	53	19.34
B385452	6.19	8.81	hypothetical protein	34	24	53	19.34
B603951	6.54	8.83	TonB-dependent siderophore receptor	35	24	62	18.22
B362201	6.29	9.13	hypothetical protein	34	26	78	16.17
B58052	13.92	10.05	EamA family transporter	34	27	86	15.62
B538146	7.08	8.12	Transposase	26	27	57	15.45
B381713	3.13	8.1	aerobactin synthetase subunit alpha	26	27	69	13.71
B381836	4	8.36	N(6)-hydroxylysine O-acetyltransferase	26	27	70	13.51
B385021	6.31	9.37	response regulator transcription factor	29	20	52	13.37

B381588	3.73	8.1	MFS transporter	26	27	71	13.34
B382081	3.66	7.95	NADPH-dependent L-lysine N(6)-monooxygenase	26	27	71	13.34
B382206	3.15	7.96	ferric aerobactin receptor IutA	26	27	71	13.34
B382654	12.91	10.43	hypothetical protein	23	27	57	13.21
B382762	12.68	10.11	DM13 domain-containing protein	23	27	57	13.21
B382870	12.57	9.89	hypothetical protein	23	27	57	13.21
B597737	-13.54	-3.82	GTP-binding protein	37	27	122	13.16
B381162	13.15	9.85	c-type lysozyme inhibitor	23	27	58	12.99
B381271	13.55	9.93	peptide deformylase	23	27	58	12.99
B382331	13.15	9.44	hypothetical protein	23	27	58	12.99
B385565	5.41	9.45	hypothetical protein	27	24	58	12.86
B385675	5.49	9.6	hypothetical protein	27	24	58	12.86
B382547	12.47	10.44	hypothetical protein	22	27	57	12.36
B382440	12.94	9.48	TetR/AcrR family transcriptional regulator	22	27	57	12.2
B381960	3.51	7.76	IucA/IucC family siderophore biosynthesis protein	25	26	69	12
B402327	13.69	9.66	Tn3 family transposase	21	27	55	11.84
B380773	13.47	10.13	alpha/beta hydrolase	23	27	72	10.76
B239784	6.83	4.01	hypothetical protein	30	27	107	10.55
B385127	6.08	10.14	putative protein	23	23	49	10.46
B402432	7.56	11.2	hypothetical protein	15	26	31	10.37
B384912	6.93	9.37	response regulator	26	19	54	9.88
B216873	6.27	4.7	putative protein	28	26	99	9.75
B383235	11.1	10.24	putative protein	18	27	55	9.22
B390200	13.43	9.18	putative transposase	20	26	65	9.05
B384119	11.63	9.66	NAD(P)-dependent alcohol dehydrogenase	18	27	55	9.01
B378656	8.49	7.31	putative partitioning ATPase protein	15	27	37	8.9
B381485	10.89	10.06	putative protein	18	27	57	8.81
B378577	8.27	6.91	putative protein	15	27	37	8.8
B383807	11.04	9.43	putative protein	18	27	58	8.72
B383910	11.49	9.9	putative transcriptional regulator	18	26	54	8.58
B378735	7.98	6.87	chromosome partitioning protein ParB	16	27	43	8.52
B383335	11.76	9.72	putative protein	18	27	59	8.46
B393423	3.8	7.5	ABC transporter ATP-binding protein	22	23	65	8.42
B378821	7.38	6.99	putative protein	16	27	43	8.36
B378907	7.6	7.12	putative protein	16	27	43	8.36

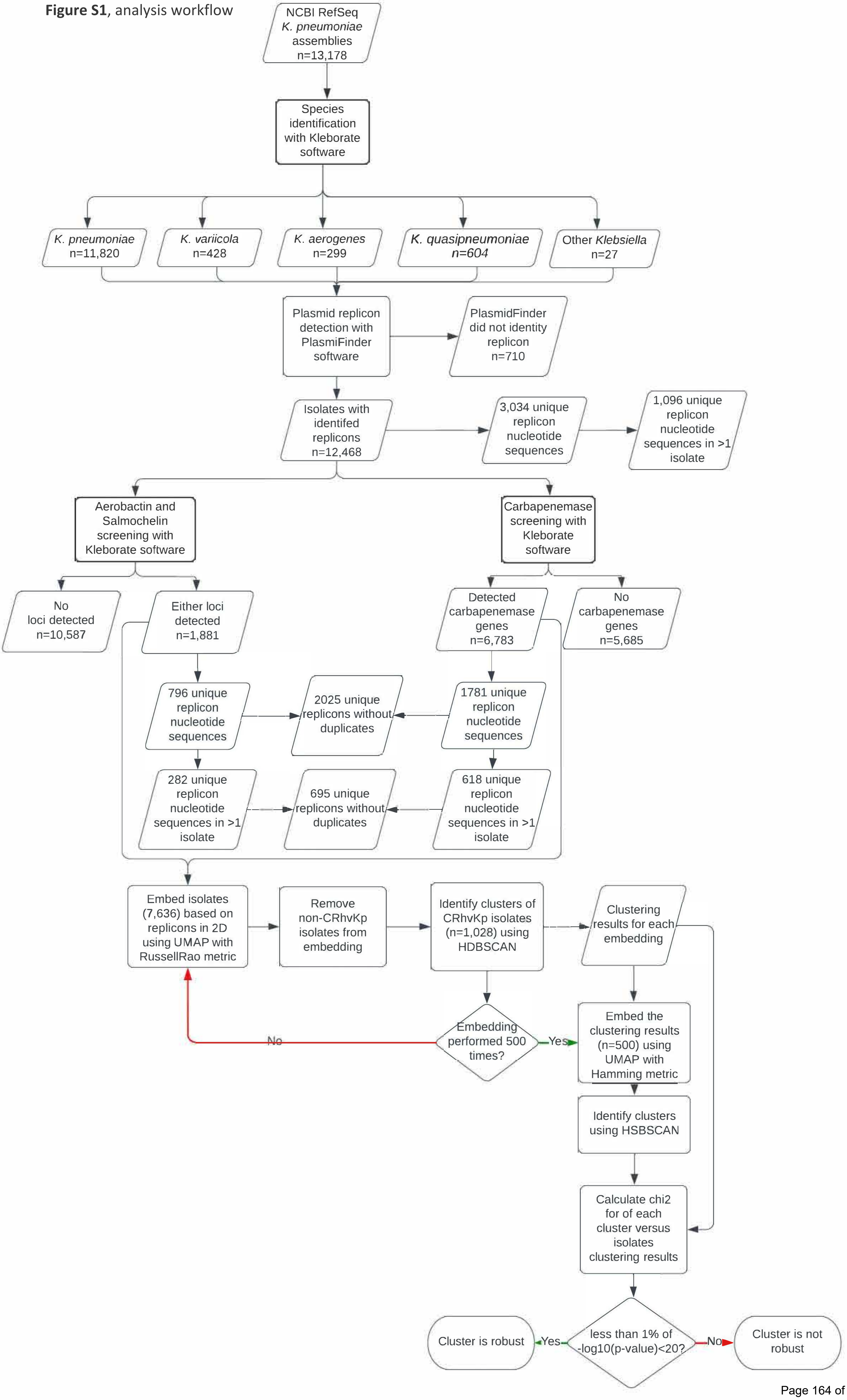
B242313	6.74	5.14	SOS mutagenesis and repair protein UmuC	23	27	85	8.25
B383439	11.65	9.36	nucleotide-binding protein	18	27	61	8.18
B388835	7.15	11.02	aldo/keto reductase	15	27	45	8.17
B378264	7.66	8.23	Cd(II)/Pb(II)-responsive transcriptional regulator	15	27	45	8.06
B385885	6.62	9.97	IS3 family transposase	19	24	46	7.97
B554821	-13.1	-4.21	Putative metal chaperone	28	24	115	7.93
B593723	7.69	4.65	hypothetical protein	24	27	97	7.79
B388922	7.4	11.41	hypothetical protein	15	27	47	7.79
B378173	7.98	8.05	heavy metal translocating P-type ATPase	15	27	49	7.63
B383121	11.2	10.41	hypothetical protein	18	27	66	7.52
B239951	6.48	3.77	putative protein	24	27	100	7.4
B378087	7.72	8.07	integral membrane protein fused with prolipoprotein signal peptidase	15	26	45	7.38
B380897	-12.94	-4.14	transcriptional repressor	26	26	116	7.32
B389471	7.66	10.96	putative N-carbamyl-L-amimo acid amidohydrolase	13	27	43	7.31
B389011	7.17	11.97	ABC transporter permease subunit	15	27	50	7.3
B389103	7.85	11.66	ABC transporter substrate-binding protein	15	27	50	7.3
B389195	6.95	11.57	MFS transporter	15	27	50	7.3
B389287	7.22	12.02	ABC transporter ATP-binding protein	15	27	50	7.3
B389379	7.62	12.09	Oligopeptide transport system permease protein OppB	15	27	50	7.3
B388716	7.36	11.96	hypothetical protein	16	26	52	7.27
B545460	6.47	10.72	IS1 transposase orfA	23	21	62	7.21
B307142	11.36	10.59	korC	21	27	95	7.21
B216664	5.25	3.61	hypothetical protein	30	26	147	7.16
B607670	11.83	8.98	transcriptional regulator	22	27	118	6.87
B151867	13.49	10.36	recombinase family protein	23	27	107	6.63
B379932	10.01	5.54	GNAT family N-acetyltransferase	13	27	42	6.2
B384219	10.66	9.68	AraC family transcriptional regulator	15	27	57	6.13
B380096	10.11	4.74	MSMEG_0570 family nitrogen starvation response protein	13	27	43	6.1
B229379	5.76	2.86	hypothetical protein	29	24	147	6.04
B379682	9.39	4.95	MSMEG_0572 family nitrogen starvation response protein	13	27	43	6.04

B379598	9.91	4.84	Transcriptional regulator%2C GntR family domain / Aspartate aminotransferase	13	27	44	5.84
B379765	9.99	5.16	Nit6803 family nitriliase	13	27	44	5.84
B393534	-13.18	-3.78	ABC transporter substrate-binding protein	24	25	113	5.83
B393737	-13.07	-3.36	iron (III) ABC transporter permease protein	22	21	91	5.8
B283147	5.89	4.34	putative protein	22	27	101	5.73
B389554	8.12	10.27	putative protein	13	27	51	5.53
B389645	8.4	10.96	putative protein	13	27	51	5.53
B389736	8.48	10.61	putative protein	13	27	51	5.53
B380179	9.56	5.51	MSMEG_0569 family flavin-dependent oxidoreductase	13	26	43	5.33
B380014	9.74	5.14	sll0787 family AIR synthase-like protein	12	27	43	5.17
B390748	7.08	5.89	transposase	18	26	80	5.02
B411899	-13.92	-5.13	cysteine--tRNA ligase	19	24	96	4.9
B395187	-13.36	-4.67	RNA polymerase-binding protein DksA	20	25	101	4.72
B241036	5.98	3.31	putative protein	24	27	142	4.71
B391152	6.08	4.9	integrase	23	26	117	4.66
B411895	-13.77	-4.94	dihydroorotase	20	26	102	4.42
B394151	-14.3	-4.96	putative protein	20	25	100	4.36
B390321	-13.66	-4.96	M14 family metallocarboxypeptidase	20	26	105	4.33
B390873	6.88	5.74	transposase	19	27	107	4.3
B242735	7.66	7.56	putative protein	15	27	74	4.28
B606549	5.74	3.89	recombinase	24	27	149	4.21
B241340	5.47	3.3	DNA replication protein	24	27	151	4.17
B241545	5.76	3.51	putative protein	24	27	151	4.17
B276809	5.33	3.11	putative protein	24	26	147	4.04
B233549	7.14	7.56	sensor domain-containing diguanylate cyclase	16	27	78	4.03
B387865	-13.16	-4.95	porphobilinogen synthase	20	25	104	4.01
B570661	5.91	3.14	putative protein	24	27	154	4
B363477	8.17	9.75	Cysteinyl-tRNA synthetase	14	21	48	3.93
B607698	5.45	3.84	putative protein	23	27	147	3.84
B602282	-14.35	-4.53	ferrous iron transporter B	19	22	97	3.83
B393874	10.64	10.03	Putative metal chaperone%2C involved in Zn homeostasis	12	20	44	3.77
B579299	11.9	10.52	transposase IS3/IS911 family protein	19	24	104	3.75
B402761	-13.58	-4.65	hypothetical protein	19	21	98	3.7
B411969	7.59	10.01	Cysteinyl-tRNA synthetase	14	13	36	3.7

B379479	9.17	5.14	putative protein	11	25	41	3.63
B68511	9.6	6.1	putative protein	26	27	208	3.45
B389827	8.48	10.5	transcriptional repressor	13	23	52	3.45
B596321	-14.4	-4.8	NAD(P)-binding domain-containing protein	19	25	104	3.18
B386855	-13.94	-4.29	hypothetical protein	20	19	103	3.12
B599529	11.18	9.36	IS110 family transposase	25	27	170	3.04
B378517	8.61	6.56	putative protein	7	25	28	2.93
B386690	-13.74	-4.4	isocitrate dehydrogenase	19	19	101	2.5
B429032	7.78	10.41	hypothetical protein	13	11	47	2.44
B242065	6.94	5.21	IS66 family insertion sequence hypothetical protein	22	27	193	2.36
B384510	8.12	11.64	DinG family ATP-dependent helicase YoaA	11	20	39	1.67
B384435	8.58	11.88	DEAD/DEAH box helicase	10	21	42	1.4
B217887	6.5	5.96	hypothetical protein	24	27	227	0.74

Appendix 4. Supplementary materials: Large scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem resistant hypervirulent strains.

Figure S1, analysis workflow


NCBI RefSeq *K. pneumoniae* assemblies n=13,178
Species identification with Kleborate software
K. pneumoniae n=11,820
K. variicola n=428
K. aerogenes n=299
K. quasipneumoniae n=604
Other *Klebsiella* n=27
Plasmid replicon detection with PlasmiFinder software
PlasmidFinder did not identity replicon n=710
Isolates with identifed replicons n=12,468
3,034 unique replicon nucleotide sequences
1,096 unique replicon nucleotide sequences in >1 isolate
Aerobactin and Salmochelin screening with Kleborate software
Carbapenemase screening with Kleborate software
No loci detected n=10,587
Either loci detected n=1,881
Detected carbapenemase genes n=6,783
No carbapenemase genes n=5,685
796 unique replicon nucleotide sequences
2025 unique replicons without duplicates
1781 unique replicon nucleotide sequences
282 unique replicon nucleotide sequences in >1 isolate
695 unique replicons without duplicates
618 unique replicon nucleotide sequences in >1 isolate
Embed isolates (7,636) based on replicons in 2D using UMAP with RussellRao metric
Remove non-CRhvKp isolates from embedding
Identify clusters of CRhvKp isolates (n=1,028) using HDBSCAN
Clustering results for each embedding
No
Embedding performed 500 times?
Yes
Embed the clustering results (n=500) using UMAP with Hamming metric
Identify clusters using HSBSCAN
Calculate chi2 for of each cluster versus isolates clustering results
Cluster is robust
Yes
less than 1% of -log10(p-value)<20?
No
Cluster is not robust