

Genomic Factors Associated with Invasiveness in *Klebsiella pneumoniae* Isolates from Hospitalised Adults in Johannesburg, South Africa

Denasha L. Reddy

denasha.reddy@wits-vida.org

University of the Witwatersrand

Courtney P. Olwagen

University of the Witwatersrand

Frank O. Williams

Imperial College London

Shama Khan

University of the Witwatersrand

Alane Izu

University of the Witwatersrand

Nicholas J. Dean

University of the Witwatersrand

Lara van der Merwe

University of the Witwatersrand

Jeannette Wadula

National Health Laboratory Service

Trusha Nana

National Health Laboratory Service

Rispah Chomba

National Health Laboratory Service

Sinenhlanhla Ndzabandzaba

National Health Laboratory Service

Lyle Murray

University of the Witwatersrand

Jacob Merika Tsitsi

University of the Witwatersrand

Jeremy Nel

University of the Witwatersrand

Arifa Parker

Stellenbosch University

Vicky Baillie

University of the Witwatersrand

Marit A.K. Hetland

Stavanger University Hospital

Elita Jauneikaite

Imperial College London

Ziyaad Dangor

University of the Witwatersrand

Shabir A. Madhi

University of the Witwatersrand

Article

Keywords:

Posted Date: September 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10871224/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background *Klebsiella pneumoniae* invasive disease (KPn-ID) is a leading cause of healthcare-associated morbidity and mortality in low- and middle-income countries. We report here on the genomic relatedness of invasive and colonising KPn isolates from adults hospitalised in Johannesburg, South Africa, and genetic factors associated with invasiveness.

Methods A multicentre, prospective hospital-based study of adults with culture-confirmed KPn from blood and cerebrospinal fluid was undertaken from May 15th, 2023 to May 14th, 2024. Adults in hospital without evidence of KPn-ID were matched to cases (2:1) and assessed for KPn gastrointestinal colonisation. Whole genome sequencing was performed on invasive and colonising KPn isolates. We estimated the relative invasiveness of KPn isolates based on sequence types (STs), capsular loci (KL), lipopolysaccharide (O) antigen types, virulence factors, and antimicrobial resistance (AMR) factors by calculating crude case-carrier ratios (CCRs). Additionally, multiple multivariable logistic regression models were used to measure the association between host characteristics and KPn genomic characteristics, with invasiveness.

Findings Whole genome sequencing was done on 508 invasive and 184 colonising KPn isolates. There was no single dominant clone amongst the invasive or colonising isolates on phylogenetic analysis. There was variability in the distribution of STs and K-loci between invasive (111 STs, 56 K-loci) and colonising isolates (87 STs, 57 K-loci). Several KPn genetic factors were associated with invasiveness in the crude CCR analyses: ST2497, ST39, KL2, KL64, O1αβ,2α, the virulence factors yersiniabactin, aerobactin, salmochelin and RmpADC, the genomic AMR markers for aminoglycosides, carbapenems, third generation cephalosporins, fluoroquinolones and sulphonamides. Nevertheless, after adjusting STs for patient clinical characteristics and virulence factors in multiple logistic regression models, only ST2497 (aOR 3.81; 95% CI: 1.64- 8.86), ST39 (aOR 4.50; 95% CI: 1.62- 12.52), ST307 (aOR 7.50; 95% CI: 1.67- 33.77), and yersiniabactin (aOR 2.97; 95% CI: 1.83- 4.84) remained significantly associated with invasiveness. After further adjustment in a separate model for K-loci, O-antigen types, AMR factors and plasmids, the only genomic factor associated with invasiveness was carbapenem resistance (aOR 3.88; 95% CI: 1.28- 11.82). Similar results were obtained when using a cluster-adjusted model. Age over 65 years (aOR 2.65; 95% CI: 1.31- 5.38) was a significant clinical characteristic associated with enhanced invasive potential.

Interpretation Age over 65 years, a host susceptibility factor, was associated with invasiveness. No genomic characteristics, other than the presence of genes encoding for carbapenem resistance, were associated with a higher potential for causing invasive disease. This poses a challenge for targeted preventive strategies against KPn-ID, including development of a vaccine, which would require multiple K- and/or O-antigens for adequate coverage.

Introduction

Approximately 1.27 million deaths were attributed to infections by bacteria exhibiting antimicrobial resistance (AMR) globally in 2019, with *Klebsiella pneumoniae* (KPn) identified as the second leading pathogen.¹ Bloodstream infections (BSIs) due to KPn associated with extended-spectrum beta-lactamases (ESBL), third-generation cephalosporin resistance from ESBL or AmpC-b-lactamase production, and carbapenem-resistant strains (CRKp) are associated with increased rates of treatment failure and high case fatality risk (CFR).² *K. pneumoniae* with ESBL production or CRKp was estimated to cause 50,000 to 100,000 deaths globally in 2019,¹ the majority of which occurred in Africa (~50,000),³ positioning CRKp as a priority pathogen by the World Health Organisation (WHO) in 2024.⁴

Asymptomatic gastrointestinal colonisation with KPn is a major reservoir for the organism, and the genomic relatedness between colonising and invasive strains in the same individual is well described.⁵⁻⁸ Risk factors for KPn gastrointestinal colonisation amongst hospitalised adults include admission to an intensive care unit (ICU), extensive invasive procedures, and comorbidities such as diabetes, alcoholic liver disease, chronic obstructive pulmonary disease, chronic kidney disease and malignancies.^{5,9} A large community-based study of Norwegian adults reported that risk factors for KPn gastrointestinal colonisation were age over 60 years, travel to Greece or Asia in the past 12 months, inflammatory bowel disease, use of proton-pump inhibitors and non-steroidal anti-inflammatory drugs in the past six months, and use of antimicrobials in the past month.¹⁰ Although the pathophysiology of gastrointestinal KPn colonisation progressing to KPn-invasive disease (KPn-ID) is unclear, hypotheses include a role for organism virulence, overgrowth of colonising organisms and/or impairment of host immunity due to underlying immunosuppressive comorbidities or therapies.^{2,5}

Limited therapeutic options for KPn-ID, particularly CRKp, necessitate the development of complementary preventive strategies such as vaccination.¹¹ Key steps in the development of a KPn vaccine include identifying populations most at risk of severe morbidity and mortality from KPn-ID, and prioritising KPn strains that would be required for optimal vaccine coverage.¹¹

The terms invasive disease potential or “invasiveness” have been used interchangeably in the context of invasive pneumococcal disease (IPD), and describes invasive *Streptococcus pneumoniae* strains that transition from nasopharyngeal colonisation to IPD by entering sterile sites such as blood and cerebrospinal fluid (CSF).¹² Studies evaluating the invasiveness of *Streptococcus pneumoniae* strains using serotype-specific case-carrier ratios (CCRs), before and after the introduction of multivalent vaccines, have demonstrated that pneumococcal invasiveness varies by serotype.¹² Similarly, the invasiveness of KPn strains based on serotyping or other genomic markers could guide development of a targeted multivalent vaccine strategy against KPn. Our study objectives were to investigate the genomic relatedness of invasive and colonising KPn isolates from adults hospitalised in Johannesburg, South Africa, and to evaluate for the relative invasiveness of KPn isolates based on sequence types (STs), capsular polysaccharide loci (KL), lipopolysaccharide loci (OL), virulence factors and AMR factors.

Results

Whole genome sequencing was done on 508 invasive¹³ and 184 colonising KPn isolates (Figure 1). The detailed genomic epidemiology of the 508 invasive isolates has been reported previously,¹³ while that of the colonising isolates is available in Supplementary Tables 1a-d. Of 660 individuals enrolled as controls to investigate for KPn colonisation, nine developed KPn-ID - of whom only three were identified to have KPn gastrointestinal colonisation on preceding single cross-sectional sampling. Amongst those three individuals, only one had matching colonising and invasive isolates (25 day interval between swab and blood culture), which belonged to ST2497 (Table 1). In the other two participants the colonising and invasive isolates differed: colonising ST1271 and invasive ST5776 (134 day interval); and colonising ST985 and invasive ST9094 (44 day interval) (Table 1).

The clinical characteristics of individuals with known outcomes, from whom invasive (n=502) and colonising (n=184) KPn isolates underwent WGS, is summarised in Table 2. Briefly, male participants comprised a higher proportion of the KPn-ID group than of the colonised control group (55.2%; 277/502 vs 22.3%; 41/184, p<0.001). Cases had a higher in-hospital CFR (49.8%; 250/502) than colonised controls (3.3%; 6/184, p<0.001). As controls had been matched to cases based on age group and hospital, there were no significant differences in groups regarding these variables.

Genetic Relatedness of invasive and colonising isolates

Overall, there were 163 STs, of which 76 were only found amongst invasive isolates, and 52 amongst colonising. The most prevalent (>50 isolates each) STs present amongst both invasive (n=508) and control (n=184) isolates were: ST2497 (18.1%; 92/508 invasive and 4.3%; 8/184 colonising), ST39 (12.0%; 61/508 and 2.7%; 5/184), ST9094 (9.4%; 48/508 and 6.5%; 12/184) and ST17 (7.1%; 36/508 and 8.2%; 15/184). We evaluated the whole-genome phylogeny of the in total 692 KPn isolates annotated by clinical status (508 invasive versus 184 colonising), hospital site (CHBAH, CMJAH, HJH), and the four most prevalent STs; Figure 2. Invasive and colonising isolates were interspersed across the phylogenetic tree, with repeated representation of several STs in both groups. The phylogenetic tree revealed a polyphyletic distribution of invasive isolates, with no single clone dominating either invasive or colonising isolates. Instead, invasive isolates were scattered across multiple branches, along with colonising isolates (Figure 2).

The hospital sites were Chris Hani Baragwanath Academic Hospital (CHBAH), Charlotte Maxeke Johannesburg Academic Hospital (CMJAH), and Helen Joseph Hospital (HJH). The STs depicted were the four most common identified in invasive and colonising isolates: ST2497, ST39, ST9094, and ST17. The clusters depicted were the five largest arbitrarily named clusters identified: Cluster_001, Cluster_004, Cluster_018, Cluster_017, and Cluster_015.

The most prevalent STs belonged to genomic clusters of highly similar strains. The genomic clusters were defined as differing by a Pathogenwatch distance <10 SNPs, and all having the same sublineage (SL), KL and OL. Of the 692 isolates, 312 (45.1%) were singletons, including 121 (38.8%) of colonising and 191 (61.2%) of invasive isolates. The remaining 380 (54.9%) isolates (n=63 colonising n=317

invasive) were distributed across 47 clusters that ranged from 2 to 86 isolates per cluster (mean = 8.3, median = 3) (Figure 3). The top five clusters amongst invasive isolates were also found amongst colonising strains and across hospital sites (Figures 2 and 3): Cluster 001- SL395 (ST2497), KL64, OL2α.1 (n=78); Cluster 004- SL17 (ST9094), KL25, OL5 (n=47); Cluster 018- SL39 (ST39), KL2, OL2α.1 (n=35); Cluster 017- SL29 (ST985), KL39, OL2α.2 (n=12); and Cluster 015- SL231 (ST5785), KL30, OL2α.2 (n=13). Several STs were found in multiple clusters or as singletons, including ST17 and ST39, indicating multiple introductions of these strains across the sites.

Genomic factors associated with invasiveness

Sequence Types (STs)

Overall, the top 13 STs (>10 isolates per ST) amongst the invasive and colonising isolates were ST2497, ST39, the novel ST9094, ST17, ST5785, ST985, ST307, ST23, ST37, ST22, ST147, ST4083, and ST66 (Figure 4a). There were 76 STs found only in invasive isolates (n= 127) (Supplementary Table 2a), all of which were of low frequency, which were not included in CCR calculations or multivariable logistic regression modelling as there were no corresponding colonising isolates to compare with. The most prevalent of the exclusively invasive STs were: ST101 (1.8%; 9/508), ST86 (1.8%; 9/508), ST607 (1.4%; 7/508), and ST25 (1.2%; 6/508), with the prevalence of the remaining STs being less than 1.0%.

A summary of the crude CCRs related to ST is available in Supplementary Table 3. The only STs with a lower 95% CI limit (95% CI-LL CCR) greater than one were ST2497 (CCR 4.61; 95% CI: 2.24- 9.51) and ST39 (CCR 4.49; 95% CI: 1.84- 10.91). On multiple multivariable logistic regression modelling, after adjusting STs for patient clinical characteristics and virulence factors (Table 3; Model 2), ST2497 (aOR 3.81; 95% CI: 1.64- 8.86), ST39 (aOR 4.50; 95% CI: 1.62- 12.52), and ST307 (aOR 7.50; 95% CI: 1.67- 33.77) were associated with invasiveness. However, after further adjustment for K-loci, O-antigen types, AMR factors and plasmids (Table 3; Model 6), there were no significant positive associations between STs and invasiveness; whilst ST22 was associated with low invasiveness (aOR 0.03; 95% CI: 0.00- 0.56).

In the cluster-adjusted multivariable logistic regression model, none of the top five clusters were associated with invasiveness after adjusting for patient characteristics, virulence and AMR factors (Table 4; Model 3); whilst Cluster 017 (SL29 [ST985], KL39, OL2α.2) was associated with less invasiveness compared with other clusters (aOR 0.25; 95% CI: 0.09- 0.71).

K-loci

The top 11 K-loci overall were KL25, KL64, KL2, KL30, KL149, KL102, KL39, KL9, KL1, KL122, and KL35, which constituted 78.7% (400/508) and 50.5% (93/184) of all invasive and colonising isolates, respectively (Figure 4b). There were 18 K-loci identified solely amongst invasive isolates (7.1%; 36/508) (Supplementary Table 2b), for which the CCRs were not analysable in the absence of detection amongst colonising isolates. The most prevalent of these exclusively invasive K-loci was KL17 (2.0%; 10/508), with the remaining K-loci having an individual prevalence of less than 1.0%.

A summary of the crude CCRs in relation to K-loci is available in Supplementary Table 3. The KLs for which 95% CI-LL CCR greater than one were KL64 (CCR 5.30; 95% CI: 2.58- 10.90) and KL2 (CCR 4.12; 95% CI: 1.91- 8.91). On multiple multivariable logistic regression modelling, after adjusting for STs, patient clinical characteristics, virulence factors, other K-loci, and O-antigen types (Table 3; Model 4), KL25 was associated with invasiveness (aOR 10.48; 95% CI: 1.61- 68.26), although this association was no longer evident after adjusting for AMR factors and plasmids (Table 3; Model 6).

O-antigen types

The top 11 O-antigen types in invasive and colonising isolates were O1αβ,2α, O1αβ,2β, O5, O2β, O3γ, O2α, O4, O13, O3αβ, O10 and O11αβ,2β, which constituted 99.6% (506/508) and 100.0% (184/184) of all invasive and colonising isolates, respectively (Figure 4c). A summary of the crude CCRs related to O-antigen types is available in Supplementary Table 3. The O-antigen type O1αβ,2α was the only O-antigen type with a 95% CI-LL CCR greater than one (CCR 2.86; 95% CI: 1.90- 4.31). Nevertheless, there was no significant association between any O-antigen types and invasiveness in the multivariable logistic regression models (Table 3).

Virulence Factors

The presence or absence of the following five virulence factors was taken into account when comparing invasive and colonising isolates: yersiniabactin, colibactin, aerobactin, salmochelin, and RmpADC. A summary of the crude CCRs related to the virulence factors is available in Supplementary Table 3. Based on 95% CI-LL CCR greater than one, virulence factors associated with invasiveness were yersiniabactin (CCR 3.91; 95% CI: 2.74- 5.56), aerobactin (CCR 3.29; 95% CI: 1.43- 7.56), salmochelin (CCR 2.49; 95% CI: 1.18- 5.24), and RmpADC (CCR 3.02; 95% CI: 1.31- 6.97). After adjusting for STs, patient clinical characteristics, virulence factors, K-loci, O-antigen types and AMR factors on multivariable logistic regression modelling, only yersiniabactin was associated with invasiveness (aOR 1.95; 95% CI: 1.07- 3.55) (Table 3; Model 5). However, this association was lost after adjusting for plasmids (Table 3; Model 6). In the cluster-adjusted model, yersiniabactin was independently associated with invasiveness after adjustment for patient clinical characteristics, virulence, and AMR factors (aOR 2.32; 95% CI: 1.43- 3.76); Table 4; Model 3.

AMR Factors

The presence or absence of the following seven genomic AMR profiles was taken into account when comparing invasive and colonising isolates: aminoglycosides, carbapenems, third-generation cephalosporins, colistin, fluoroquinolones, penicillins, and sulfonamides. A summary of the crude CCRs related to the AMR factors is available in Supplementary Table 3. Five of the seven genomic AMR profiles had a 95% CI-LL CCR greater than one: aminoglycosides (CCR 2.74; 95% CI: 1.92- 3.91), carbapenems (CCR 3.56; 95% CI: 2.49- 5.09), third-generation cephalosporins (CCR 3.14; 95% CI: 2.22- 4.46), fluoroquinolones (CCR 3.14; 95% CI: 2.22- 4.46), and sulfonamides (CCR 2.80; 95% CI: 1.97- 3.98). On multivariable logistic regression modelling, after adjusting for STs, patient clinical characteristics,

virulence factors, K-loci, O-antigen types and plasmids, there was a significant association between the presence of genes encoding for carbapenem resistance and invasiveness (aOR 3.88; 95% CI: 1.28- 11.82); Table 3; Model 6. The same independent association was seen in the cluster-adjusted model (aOR 2.34; 95% CI: 1.23- 4.46); Table 4.

Plasmids

The presence or absence of 14 plasmids was taken into account when comparing invasive and colonising isolates (Supplementary Table 4). The most prevalent plasmids (>100) were IncFIB (68.7%; 173/252 invasive vs 31.3%; 79/252 colonising), IncFIB/IncFII (73.0%; 170/233 invasive vs 27.0%; 63/233 colonising), IncX3 (78.1%; 132/169 invasive vs 21.9%; 37/169 colonising), IncFIB/IncHI1B (92.4%; 121/131 invasive vs 7.6%; 10/131 colonising), and IncR (65.2%; 73/112 invasive vs 34.8%; 39/112 colonising); Supplementary Table 4. A summary of the crude CCRs related to the listed plasmids is available in Supplementary Table 3. Of the 14 plasmids included in the analysis, two had a 95% CI-LL CCR greater than one: IncFIB/IncFII/IncR (CCR 2.67; 95% CI: 1.08- 6.63) and IncFIB/IncHI1B (CCR 5.21; 95% CI: 2.71- 10.03), which was no longer evident following multiple logistic regression modelling (Table 3; Model 6).

Clinical characteristics associated with invasiveness

On multivariable logistic regression modelling, age over 65 years (aOR 2.65; 95% CI: 1.31- 5.38) was significantly associated with invasiveness, after adjustment for STs, hospital site, virulence factors, K-loci, O-antigen types, AMR factors, and plasmids, respectively (Table 3; Model 6). This association also remained in the cluster-adjusted models (aOR 2.27; 95% CI: 1.19- 4.32); Table 4. The hospital site, CMJAH, was also independently associated with invasiveness after adjustment for all genomic factors (aOR 1.61; 95% CI: 1.02- 2.55) (Table 3; Model 6) and after cluster-adjustment (aOR 1.70; 95% CI: 1.12- 2.59) (Table 4).

Discussion

Our study used crude CCRs and multiple multivariable logistic regression models to identify KPn genomic factors associated with enhanced invasiveness in isolates from a large prospective South African cohort. A key finding was that although several KPn genomic factors were associated with invasiveness (versus colonisation) in crude CCR analyses, we did not find any STs, K-loci, O-antigen types, or virulence factors to be independently associated with invasiveness in multiple multivariable logistic regression models, following adjustment for genomic variables (including AMR factors and plasmids) and clinical characteristics. Notably, KPn isolates with genes encoding for carbapenem resistance, and the clinical characteristic of age over 65 years remained significantly associated with invasiveness in all models tested.

No specific STs were significantly associated with invasiveness in our regression models after adjustment for K-loci, O-antigen types, genomic AMR factors and plasmids. This finding is in contrast to

other pathogens such as *Streptococcus pneumoniae* and Group B *Streptococcus* where differences of invasiveness have been described based on the serotype-specific polysaccharide capsules.^{12,14} A possible explanation is that IPD is a community-acquired infection driven by organism virulence factors such as the polysaccharide capsule and host susceptibility factors.¹² In contrast, KPn-ID is predominantly a healthcare-associated infection perpetuated by AMR-related hospital outbreaks, in addition to host risk factors.⁵ Further, our findings of genetic similarity between invasive and colonising isolates indirectly support other studies that have reported that most patients with KPn-ID develop invasive disease from their own gastrointestinal colonising strain.^{6,15}

In our model, the lack of association between specific STs, K-loci and O-antigen types with enhanced invasiveness in KPn poses a potential challenge with regard to development of a multivalent vaccine that would aim to target the most prevalent strains causing KPn-ID in adults. Although it has been estimated that 15-20 K-antigen types were responsible for >70.0% of global KPn-ID in infants, their diversity poses a challenge to developing a multivalent vaccine against K-antigens.¹⁶⁻¹⁸ A glycoconjugate vaccine targeting a few selected O-antigens could protect against most strains of KPn causing invasive disease worldwide,^{19,20} however the capsular polysaccharide has been shown to inhibit and interfere with O-antigen antibody binding and functioning, which could limit the effectiveness of a vaccine targeting O-antigens.²¹

The association of age over 65 years with invasiveness is expected as the correlation between older age and morbidity from KPn-ID is well described,²² particularly in the context of CRKp.^{23,24} One study site, CMJAH, was independently associated with invasiveness in all adjusted models. This is likely related to unmeasured confounders in the analysis including that CMJAH is a quaternary level facility with the highest number of ICUs amongst the study sites sampled, which is in itself associated with higher risk for KPn-ID.²² As previously reported by our study group, most cases with KPn-ID in our cohort (40.8%) had been admitted to ICU.²⁵

In contrast to a Malawian study²⁶ where the siderophore virulence factor yersiniabactin was more common in colonising isolates, we noted yersiniabactin to have a crude CCR greater than one, and to be independently associated with invasiveness in the cluster-adjusted model and in the primary logistic regression analysis prior to adjustment for plasmids. The loss of yersiniabactin's association with invasiveness following adjustment for plasmids was an unexpected finding in our study, and should be explored in other KPn invasiveness analyses. The association between mobile integrative conjugative element (ICEKp)-mediated yersiniabactin and KPn-ID has been well described,²⁷ particularly in the context of yersiniabactin and CRKp isolates.²⁸⁻³⁰ A hypothesis for the link between yersiniabactin and CRKp involves structural similarities between the KPn siderophore importer protein YbtPQ and antibiotic efflux pumps, with improved antibiotic susceptibility following deletion of ICEKp demonstrated in vivo.³¹ Further, as ICEKp is highly mobile, this "secondary" mechanism for AMR can be transmissible within CRKp outbreak strains.³¹

Regarding AMR factors, the significant association between genomic carbapenem resistance and invasiveness was likely related to unmeasured clinical confounders that drive KPn-ID, such as exposure to antimicrobials, ICU admission, invasive procedures, and prolonged hospital stays,³² which were not included in the logistic regression models due to missing clinical data. Thus, a limitation of our study included not having detailed clinical information for all cases of KPn-ID, although we attempted to minimise sampling bias and provide accurate molecular epidemiological data by analysing all available invasive KPn isolates. Another limitation was that plasmids were analysed using short reads, any further and more detailed analyses should be done using long reads nanopore or PacBio technology. Another limitation was that the cluster analysis was not linked with temporal or epidemiological data, and thus transmission dynamics could not be analysed (although that was not the focus of this study). However, the use of a cluster-adjusted model allowed adjustment of genomic factors that were part of clonal expansions, which could otherwise lead to an inflated association with invasiveness.

As the aim of our study was to compare the genomic epidemiology of invasive and colonising isolates, we did not evaluate for gastrointestinal colonisation in the cases, but rather assumed that all cases would have been colonised. Thus, we could not confirm that cases in our cohort developed invasive disease from their own gastrointestinal colonising strain, although this has been investigated in other studies designed to test this hypothesis.^{6,15} Further, gastrointestinal colonisation is dynamic and our study included a single rectal swab at enrolment,³³ thus not accounting for changes in colonising KPn populations over time and/ or the presence of multiple colonising KPn strains simultaneously, which may explain the discordant colonising and invasive strains isolated from two patients in our study. Future studies investigating KPn strain invasiveness may benefit from employing a prospective cohort study design with multiple timepoints for rectal sampling and use of genomic methods that allow for detection of multiple concurrent KPn strains, to address important unanswered questions regarding colonisation dynamics and risk factors for invasiveness. Of particular interest would be whether patients' colonising isolates are acquired in hospital or represent the patient's own microbiota; whether most individuals transition from an uncolonised to colonised state prior to developing invasive disease; and whether wild-type strains (if isolated initially) remain present throughout admission in colonised individuals or are replaced by hospital-acquired strains.

In conclusion, based on crude CCRs and multiple multivariable logistic regression models, we did not identify any significant differences in serotype, KL or O-antigen invasiveness between invasive and colonising isolates. An implication of this finding is that no specific genomic factors inform vaccine inclusion based on invasiveness, and that a KPn vaccine for adults should instead target the most common invasive strains, which will likely require a multivalent approach given the diversity in KL. As previous studies have demonstrated the overlap between colonising and invasive strains in patients with KPn-ID, in the absence of invasive isolates, colonisation studies may provide an indirect measure of most common invasive isolates. The host susceptibility factor of age over 65 years was associated with invasiveness and this population group would benefit from KPn-ID preventive strategies, including vaccination.

Methods

Study design and study population

The study was conducted prospectively from May 15th, 2023 to May 14th, 2024. Invasive disease was defined as KPn isolated from blood or CSF, identified through daily laboratory-based surveillance at the National Health Laboratory Service (NHLS) microbiology laboratories serving the three academic hospital study sites in Johannesburg, South Africa, as previously reported²⁵: Chris Hani Baragwanath Academic Hospital (CHBAH), Charlotte Maxeke Johannesburg Academic Hospital (CMJAH) and Helen Joseph Hospital (HJH). The inclusion and exclusion criteria, as well as study definitions have also been reported previously.^{13,25,33}

For each case (adults with KPn-ID), two hospitalised adult controls without KPn invasive disease were enrolled, matched to cases on age category (18-35 years of age, 36-50 years of age, 51-65 years of age, >65 years of age) and duration of admission (<one week in hospital, >one week in hospital). A single rectal swab was obtained from controls at the time of enrolment to assess for KPn gastrointestinal colonisation.³³ Cases with KPn-ID were not investigated for colonisation as we premised gastrointestinal colonisation to be a pre-requisite for developing KPn-ID.^{6-8,34} Individuals enrolled as controls, but who subsequently developed KPn-ID during hospital admission, were considered as cases. The only exclusion criterion in the study was refusal of consent to participate. Clinical and laboratory information was abstracted from clinical records and entered into a password-protected electronic database (REDCap). Participants were followed up until discharge from hospital or in-hospital death. The findings on the clinical and microbiological epidemiology of KPn-ID of the cases included in the study have been previously reported,²⁵ as well as clinical factors associated with KPn gastrointestinal colonisation,³³ and the genomic epidemiology of KPn-ID.¹³

Microbiological identification

All participating NHLS microbiology laboratories are accredited by the South African National Accreditation System. Culture, identification and antimicrobial susceptibility testing (AST) of KPn invasive isolates were carried out at the NHLS laboratories according to standard operating procedures at each site. Briefly, KPn was detected using either the Analytical Profile Index 20E (API 20E, bioMérieux, France), the VITEK 2 automated system (bioMérieux, France) or the MALDI-TOF VITEK MS (bioMérieux, France) system.

The available blood and CSF KPn isolates from eligible participants were collected from the NHLS laboratories and underwent confirmatory species testing at the Wits Vaccines and Infectious Diseases Analytics Unit (Wits-VIDA) laboratory. Briefly, single bacterial colonies were subcultured onto Colorex Orientation Agar plates (MediaMAGE, South Africa) and then incubated at 37°C for 24 hours. Following growth, the identification of KPn isolates was performed by classical microbiology techniques and confirmed by using the Analytical Profile Index 20E (API 20E, bioMérieux, France). Confirmed KPn

colonies were subcultured onto 5% Columbia Sheep Blood Agar (MediaMAGE, South Africa) to ensure purity and adequate biomass before being stored at -80°C in a medium containing skim milk, tryptone, glucose and 10% glycerol (STGG) for further analysis.

Rectal swabs, performed on all controls at enrolment to determine gastrointestinal colonisation with KPn, were inoculated into Amies transport medium (Qiagen, South Africa) at the time of collection and transported to the Wits-VIDA laboratory where traditional microbiology culture methods were performed for identification of KPn. At the Wits-VIDA laboratory, rectal swab samples were transferred to brain heart infusion (BHI) broth (Oxoid, South Africa) prior to being vortexed for 60 seconds to release microorganisms into the nutrient broth. The swabs were then incubated in the BHI broth for 24 hours at 37°C. Following enrichment, 50 µl of the BHI broth was inoculated onto Colorex Orientation Agar plates (Media Mage, South Africa) and then incubated at 37°C for up to 48 hours. Following growth, colonies were identified by morphology, as well as a chromogenic colour change, and presumptive KPn identification was confirmed on isolates with API (bioMérieux, France) 20E.

Whole genome sequencing

To prepare for whole genome sequencing (WGS), genomic DNA was extracted directly from stored STGG, using the TANBead Maelstrom 4810 Automated DNA/RNA Purification Platform (Taiwan Advanced Nanotech Inc.) following the manufacturer's guidelines. Thirty isolates required an additional pre-lysis step to meet the minimum DNA input requirements. Briefly, 50 µL of the pure culture isolate in STGG was mixed with 200 µL of lysis buffer (Taiwan Advanced Nanotech Inc.), 10 µL of proteinase K, 10 µL of lysozyme, and 10 µL of mutanolysin (M9901, Merck KGaA, Germany), followed by a staged incubation at 37 °C, 56 °C, and 60 °C before proceeding with the standard TANBead DNA extraction. Following DNA extraction, sequencing libraries were prepared using the Nextera XT DNA Kit (Illumina, US) and sequenced on the Illumina Nextseq 1000 and Illumina Nextseq 2000 platforms to generate short 2 x 150bp paired-end reads (Illumina, San Diego, CA) at the National Institute for Communicable Diseases (NICD), Johannesburg, South Africa. Quality control assessment of raw sequencing reads was performed at Imperial College London, with FastQC (version 0.12.1, 2023)³⁵ and MultiQC (version 1.21, 2024)³⁶ files, generated for raw reads, to assess initial quality metrics. Raw reads were then trimmed, with low quality reads removed using Trimmomatic (version 0.39, 2021).³⁷ This used the following parameters: a quality (Q)-score threshold of 20 for both leading and trailing bases, a minimum length for retaining informative reads of 36 bases, and a sliding window of 4 bases with an average quality threshold of 20. After trimming raw reads, Seqkit (version 2.8.2, 2024 <https://bioinf.shenwei.me/seqkit/>)³⁸ was used to estimate sequencing coverage by calculating the total number of bases relative to the reference genome size. Genome assemblies were generated from the trimmed reads using SPAdes (version 3.15.5, 2022)³⁹. A multi-k-mer approach, with k-mers of size '45,55,67,79,85,91,97,103,115,127', was used to address limitations of using a single k-mer size. Draft genomes from SPAdes were run through Quast (version 5.2.0, 2022)⁴⁰ to summarise statistics for genome assemblies. CheckM2 (version 1.0.2, 2023)⁴¹ was used to assess the completeness and

contamination of generated draft assemblies. Bioawk (version 1.0, 2015)⁴² was then utilised to remove contigs less than 200bp, with Quast⁴⁰ and CheckM2⁴¹ re-run on the filtered assembly files.

Genomic characterisation

Pathogenwatch (<https://rc.pathogen.watch/en>)⁴³ was used for species identification and KPn genomic characterisation using Kleborate v3.2.4.^{44,45} Multi-locus sequence typing (MLST) was performed using Klebsiella Pasteur MLST31 hosted on PubMLST⁴⁶ and BIGSdb-Pasteur⁴⁷, and Life Identification Number (LIN)-codes were assigned based on the core genome MLST (cgMLST) profiles to assign sublineages (SLs). Capsule K- and O- serotype predictions were made with Kaptive v3.0.⁴⁸ Antimicrobial resistance genes identified with Kleborate⁴⁴ were compared to the curated version (v3.2.4) of the Comprehensive Antibiotic Resistance Database program including ResFinder.⁴⁹ Virulence genes were identified by Kleborate⁴⁴ using BLASTn for key loci. Reconstruction of plasmids was done using MOB-suite version 3.1.8.⁵⁰ Prior to plasmid reconstruction, bacterial species from raw reads were verified using Kraken2 (version 2.1.3, 2023)⁵¹ and Bracken (version 2.9, 2023).⁵² Identification of AMR genes carried on plasmids was performed out using ABRicate (v1.0.0, 2020)⁵³ with the built-in NCBI database.⁵⁴ A reference-free, distance-based phylogenetic tree was generated from de novo genome assemblies using Split Kmer Analysis (SKA v2.0)⁵⁵ with the ska tree function. The resulting Newick-formatted tree was visualised and annotated using iTOL v7.0 (<https://itol.embl.de>).⁵⁶ Metadata including hospitals/site, invasive versus colonising classification, ST, and clusters were formatted as iTOL annotation datasets according to the iTOL input specifications and uploaded for visualisation and annotation of the phylogeny. Genomic clusters of KPn isolates were defined as genomes sharing SL, K, and O-loci, and differing by <10 Pathogenwatch SNPs, corresponding approximately to 25 genome-wide SNPs, a threshold previously shown to be indicative of hospital transmission.^{6,57-59}

Statistical Analysis

Continuous variables were described using medians and interquartile ranges (IQRs). Categorical variables were summarised using frequencies and percentages. The prevalence and 95% confidence intervals (CI) of each ST, K-locus, O-antigen type, virulence factor and AMR factor were reported by case status. Case-carrier ratios (CCRs) were estimated for ST, K-locus, O-antigen type, virulence factor and AMR factor, and presented alongside 95% CIs.⁶⁰ We defined an association with heightened invasiveness as a lower bound of the 95% confidence interval of the CCR of more than one. The CCRs were only calculated for KPn-specific STs, K-loci, or O-antigens represented among invasive isolates and also detected among colonising isolates. A Wald test was used to compare odds of characteristic in cases and controls, and p-values of <0.05 were considered statistically significant. The equation used was:

$$\text{CCR} = ((\text{number of cases for variable} + 0.5) * (\text{number of controls not for variable} + 0.5)) / ((\text{number of controls for variable} + 0.5) * (\text{number of cases not for variable} + 0.5))$$

Multiple multivariable logistic regression models were used to investigate the association of genomic variables with invasiveness in a group of 508 cases and 184 controls. The top 13 KPn STs (>10 isolates represented in each ST) were included as a categorical model in all models, as well as patient clinical characteristics (age category and study site), which were considered potential confounders for invasiveness. The following groups of genomic variables were sequentially added to each logistic regression model: virulence-encoding genes *yersiniabactin*, *colibactin*, *aerobactin*, *salmonella* and *rmpADC*; the top ten K-loci; the top nine O-antigen types; genomic AMR profiles; and the top 14 plasmids identified. Adjusted odd ratios and the associated 95% CIs were estimated using multivariable logistic regression analysis. A separate cluster-adjusted multivariable regression model was used, using the top five genomic clusters as baseline categorical variables to assess if observed associations with invasiveness were driven by putative transmission-linked genomic clusters or if genomic associations with invasiveness remained independent of cluster structure. Data was analysed with STATA Version 18.5 (StataCorp, Texas, USA) and R Version 4.1.1 (Vienna, Austria).

Declarations

Ethical approval

The study was approved by the Human Research Ethics Committee (HREC), University of the Witwatersrand (HREC reference number: M220960), and registered on the South African National Health Research Database (reference number: GP202209026). Controls were only enrolled for the cases that were enrolled alive. Written informed consent was obtained for individuals enrolled during their hospital admission. If individuals were too ill to provide consent at the time of enrolment, proxy consent was obtained from their next-of-kin. Ethical waivers were granted by the HREC for enrolment of cases where an individual was discharged or died before consent could be obtained, to enable retrospective enrolment of the cases for whom medical records could be retrieved, and for the collection of isolates, but not clinical information, from individuals with KPn-ID who were excluded from the clinical study. Genomic analysis was performed on all KPn isolates collected during the study.

Data availability

All sequencing reads generated in this study have been deposited in the NCBI genome database under the BioProjects PRJNA1369306 and PRJNA1219545. The complete list of accession numbers for BioProjects PRJNA1369306 (n=507 invasive genomes and n=184 colonising genomes) and PRJNA1219545 (n=1 invasive genome KPIDA0157) has been added to the excel supplementary file.

Supplementary Data

Supplementary materials are available online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Author contributions. D.L.R, Z.D and S.A.M conceptualised and designed the study. D.L.R collected and analysed the clinical and microbiological data, analysed the molecular data, and wrote the first draft of the manuscript. A.I performed the CCR statistical analyses, and D.L.R performed the multiple logistic regression models. N.J.D, L.V.D.M, V.L.B and C.P.O facilitated the molecular testing (DNA extractions and whole genome sequencing) of isolates. S.K performed the bioinformatics analysis. E.J and F.W performed genomic quality control, reconstruction of plasmids, characterisation and analysis. M.A.K.H performed the genomic cluster identification. S.A.M, C.P.O, S.K, N.J.D, L.V.D.M, Z.D, J.W, T.N, R.C, S.N, E.J, F.O.W, M.A.K.H, A.P, A.I and V.B contributed to the final manuscript.

Acknowledgments. The authors thank the staff at the NHLS laboratories and the laboratory team at Vaccines and Infectious Diseases Analytics Research Unit (Wits-VIDA). We also thank the Institut Pasteur teams for the curation and maintenance of BIGSdb-Pasteur databases at <http://bigsdb.pasteur.fr>. E.J acknowledges funding from Genesis Research Trust. E.J and F.W acknowledge computational resources and support provided by the Imperial College Research Computing Services.

Financial Support. This study was funded internally by Wits-VIDA, Johannesburg, South Africa. D.L.R acknowledges funding from the South African Medical Research Council (SAMRC) through its Division of Research Capacity Development under the Clinician Researcher Development Programme from funding received from the Public Health Enhancement Fund / South African National Department of Health 2023. The content hereof is the sole responsibility of the authors and does not necessarily represent the official views of the SAMRC. D.L.R. acknowledges funding from the Netcare Physicians Partnership Trust (Professor Bongani Mayosi Netcare Clinical Scholarship).

Potential Conflicts of Interest. All authors report no potential conflict of interests.

References

1. Murray, C. J. L. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
2. Chang, D. et al. Risk Factors, and Control Strategies of *Klebsiella pneumoniae* Infection. *Front Microbiol* ; 12. (2021).
3. Sartorius, B. et al. The burden of bacterial antimicrobial resistance in the WHO African region in 2019: a cross-country systematic analysis. *Lancet Glob Health* ; 12. (2024).
4. WHO Bacterial Priority Pathogens List. 2024. (2024).
<https://www.who.int/publications/i/item/9789240093461> [Accessed 2 March 2025].
5. Wyres, K. L. & Lam, M. M. C. Holt. K. Population genomics of *Klebsiella pneumoniae*. *Nat. Rev. Microbiol.* **18**, 344–359 (2020).

6. Gorrie, C. L. et al. Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients. *Clin. Infect. Dis.* **65** (2), 208–215 (2017).
7. Gorrie, C. L. et al. Antimicrobial-Resistant *Klebsiella pneumoniae* Carriage and Infection in Specialized Geriatric Care Wards Linked to Acquisition in the Referring Hospital. *Clin. Infect. Dis.* **67** (2), 161–170 (2018).
8. Martin, R. M. et al. Molecular Epidemiology of Colonizing and Infecting Isolates of *Klebsiella pneumoniae*. *mSphere* (2016).
9. Perovic, O. et al. Carbapenem-resistant Enterobacteriaceae in patients with bacteraemia at tertiary hospitals in South Africa, 2015 to 2018. *Eur. J. Clin. Microbiol. Infect. Dis.* **39** (7), 1287–1294 (2020).
10. Raffelsberger, N. et al. Gastrointestinal carriage of *Klebsiella pneumoniae* in a general adult population: a cross-sectional study of risk factors and bacterial genomic diversity. *Gut Microbes* ; **13**(1). (2021).
11. Dangor, Z. et al. Vaccine value profile for *Klebsiella pneumoniae*. *Vaccine* (2024).
12. Gallagher, K. E. et al. Serotype-specific pneumococcal invasiveness: a global meta-analysis of paired estimates of disease incidence and carriage prevalence. *Lancet Microbe.* **7** (3), 101301 (2026).
13. Reddy, D. L. et al. *Preprint. Population genomics, resistome and plasmidome of invasive Klebsiella pneumoniae isolates in adults hospitalised in Johannesburg, South Africa* (Research Square, 2026).
14. Bianchi-Jassir, F. et al. Systematic review of Group B Streptococcal capsular types, sequence types and surface proteins as potential vaccine candidates. *Vaccine* **38** (43), 6682–6694 (2020).
15. Martin, R. M. et al. Molecular Epidemiology of Colonizing and Infecting Isolates of *Klebsiella pneumoniae*. *mSphere* ; **1**(5). (2016).
16. Wantuch, P. L. & Rosen, D. *Klebsiella pneumoniae*: adaptive immune landscapes and vaccine horizons. *Trends Immunology* ; **44**(10). (2023).
17. Wyres, K. L. et al. Genomic surveillance for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from South and Southeast Asia. *Genome Medicine* ; **12**(11). (2020).
18. Stanton, T. D. et al. Distribution of capsule and O types in *Klebsiella pneumoniae* causing neonatal sepsis in Africa and South Asia: A meta-analysis of genome-predicted serotype prevalence to inform potential vaccine coverage. *PLoS Med.* **Jan 12** (1), e1004879 (2026).
19. Choi, M. et al. The Diversity of Lipopolysaccharide (O) and Capsular Polysaccharide (K) Antigens of Invasive *Klebsiella pneumoniae* in a Multi-Country Collection. *Front Microbiol* (2020).
20. Miller, J. C., Cross, A. S., Tennant, S. M. & Baliban, S. *Klebsiella pneumoniae* Lipopolysaccharide as a Vaccine Target and the Role of Antibodies in Protection from Disease. *Vaccines* **12**, 1177 (2024).
21. Hwang, W. et al. Antibody responses in *Klebsiella pneumoniae* bloodstream infection: a prospective cohort study. *Lancet Microbe.* **6**, 100988 (2025).
22. Roach, D. J. et al. Clinical and Genomic Characterization of a Cohort of Patients With *Klebsiella pneumoniae* Bloodstream Infection. *Clin. Infect. Dis.* **78** (1), 31–39 (2024).

23. Chen, Y. et al. Risk factors and mortality for elderly patients with bloodstream infection of carbapenem resistance *Klebsiella pneumoniae*: a 10-year longitudinal study. *BMC Geriatr.* **22**, 573 (2022).
24. Colkesen, F. et al. Carbapenem-Resistant *Klebsiella pneumoniae* Infection and Its Risk Factors in Older Adult Patients. *Clinical Interventions in Aging* ; 18;1037–1045. (2023).
25. Reddy, D. L. et al. Clinical and microbiological epidemiology of *Klebsiella pneumoniae* invasive disease in hospitalized adults in Johannesburg, South Africa: a multicentre observational study. *JAC Antimicrob. Resist* ; **8**(1). (2026).
26. Lewis, J. M. et al. Genomic and antigenic diversity of colonizing *Klebsiella pneumoniae* isolates mirrors that of invasive isolates in Blantyre, Malawi. *Microbial Genomics* (2022).
27. Lam, M. M. C. et al. Genetic diversity, mobilisation and spread of the yersiniabactin encoding mobile element ICEKp in *Klebsiella pneumoniae* populations. *Microbial Genomics* (2018).
28. Cienfuegos-Gallet, A. V., Zhou, Y., Ai, W., Kreiswirth, B. N. & Yu, F. Chen. L. Multicenter Genomic Analysis of Carbapenem-Resistant *Klebsiella pneumoniae* from Bacteremia in China. *Microbiology Spectrum* (2022).
29. Jati, A. P. et al. Widespread Detection of Yersiniabactin Gene Cluster and Its Encoding Integrative Conjugative Elements (ICEKp) among Nonoutbreak OXA-48-Producing *Klebsiella pneumoniae* Clinical Isolates from Spain and the Netherlands. *Microbiology Spectrum* ; **11**(4). (2023).
30. On, Y., Kim, J. W., Lee, J. & Yoo, J. Genomic analysis of carbapenem-resistant *Klebsiella pneumoniae* blood isolates from nationwide surveillance in South Korea. *Front. Microbiol.* **16**, 1562222 (2025).
31. Farzand, R. et al. A Virulence Associated Siderophore Importer Reduces Antimicrobial Susceptibility of *Klebsiella pneumoniae*. *Front. Microbiol.* **12**, 607512 (2021).
32. Zhu, Q., Xu, J., Chen, X., Ren, Y. & Zhao, L. Risk factors and molecular epidemiology of bloodstream infections due to carbapenem-resistant *Klebsiella pneumoniae*. *Diagnostic Microbiol. Infect. Disease* (2023).
33. Reddy, D. L. et al. Factors associated with *Klebsiella pneumoniae* colonisation in adults hospitalised in Johannesburg, South Africa. *S Afr. J. Infect. Dis.* **41** (1), a804 (2026).
34. Martin, R. M. & Colonization, B. M. Infection, and the Accessory Genome of *Klebsiella pneumoniae*. *Frontiers Cell. Infect. Microbiology* ; 8. (2018).
35. Andrews, S. FastQC. A quality control tool for high throughput sequence data. (2023).
<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
36. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32** (19), 3047–3048 (2016).
37. Bolger, A. M. & Lohse, M. Usadel. B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30** (15), 2114–2120 (2014).
38. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss Army Knife for Sequence and Alignment Processing. *iMeta* (2024).

39. Bankevich, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal Comput. Biology* (2012).
40. Gurevich, A., Saveliev, V. & Vyahhi, N. Tesler. G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29** (8), 1072–1075 (2013).
41. Chklovski, A. CM2 Reference Database Uniref100/KO (Version 3) [Data set]. Zenodo; (2025).
42. BWK awk modified for biological data. January (2026). <https://github.com/lh3/bioawk?tab=readme-ov-file> (accessed 30).
43. Argimon, S. et al. Rapid Genomic Characterization and Global Surveillance of *Klebsiella* Using Pathogenwatch. *Clinical Infect. Diseases* ; (73). (2021).
44. Lam, M. M. C. et al. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nature Communications* ; 12. (2021).
45. Whitfield, C. et al. O-antigen polysaccharides in *Klebsiella pneumoniae*: structures and molecular basis for antigenic diversity. *Microbiology Mol. Biology Reviews* ; **89**(2). (2025).
46. PubMLST. Public databases for molecular typing and microbial genome diversity. (2026).
47. Insitut Pasteur BIGSdb-Pasteur. (2026).
48. Lam, M. M. C., Wick, R. R., Judd, L. M. & Holt, K. E. Wyres. K. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the *Klebsiella pneumoniae* species complex. *Microbial Genomics* ; 8. (2022).
49. Florensa, A. F., Kaas, R. S., Clausen, P. T. L. C., Aytan-Aktug, D. & Aarestrup F. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microbial Genomics* ; 8. (2022).
50. Robertson, J. & Nash, J. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* **4**, e000206 (2018).
51. Wood, D. E. & Lu, J. Langmead. B. Improved metagenomic analysis with Kraken 2 (2019). *Genome Biology* (2019).
52. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg S. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Science* (2017).
53. ABRicate. (2026).
54. Feldgarden, M. et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype Phenotype Correlations in a Collection of Isolates. *Antimicrobial Agents Chemotherapy* ; **63**(11). (2019).
55. Harris, S. Preprint. SKA: Split Kmer Analysis Toolkit for Bacterial Genomic Epidemiology. (2018).
56. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* (2021).
57. Foster-Nyarko, E. et al. Nanopore-only assemblies for genomic surveillance of the global priority drug-resistant pathogen, *Klebsiella pneumoniae*. *Microbial Genomics* ; 9. (2023).

58. Sherry, N. L. et al. Genomics for Molecular Epidemiology and Detecting Transmission of Carbapenemase-Producing Enterobacterales in Victoria, Australia, 2012 to 2016. *J Clin. Microbiol* (2019).

59. David, S. et al. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat Microbiol* (2019).

60. Smith, T., Lehmann, D., Montgomery, J., Grattens, M. & Riley, I. D. Alpers. M. Acquisition and invasiveness of different serotypes of *Streptococcus pneumoniae* in young children. *Epidemiol. Infect.* **111**, 27–39 (1993).

Tables

Tables 1 to 4 are available in the Supplementary Files section.

Figures

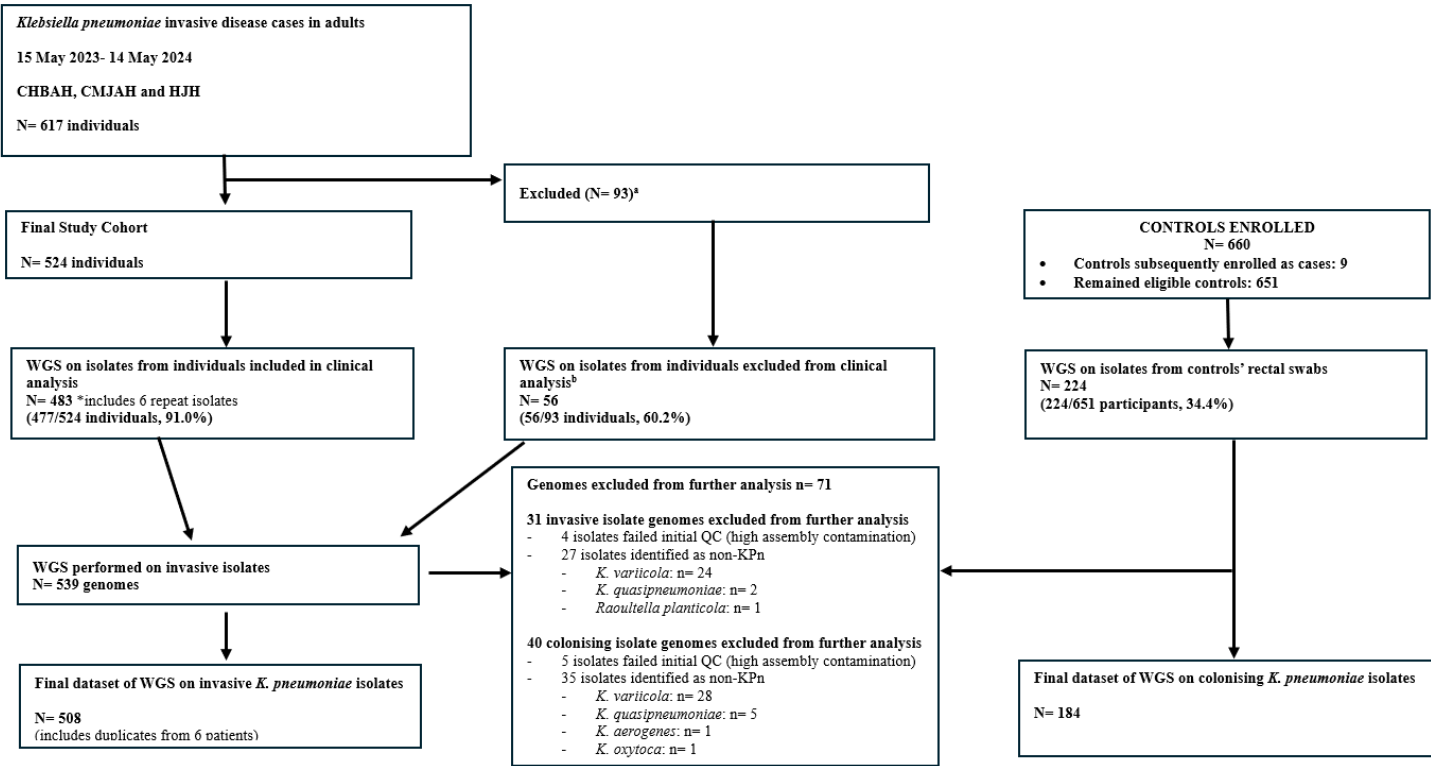


Figure 1

Consort diagram

^a 93 cases were excluded from clinical analysis for the following reasons: records could not be traced (n=40), declined participation in study (n=33), discharged prior to enrolment before ethical approval granted for retrospective enrolment (n=11), and missed on initial laboratory surveillance (n=9).

^b An ethical waiver was granted for the collection of isolates, but not clinical information, from excluded individuals with KPn-ID.

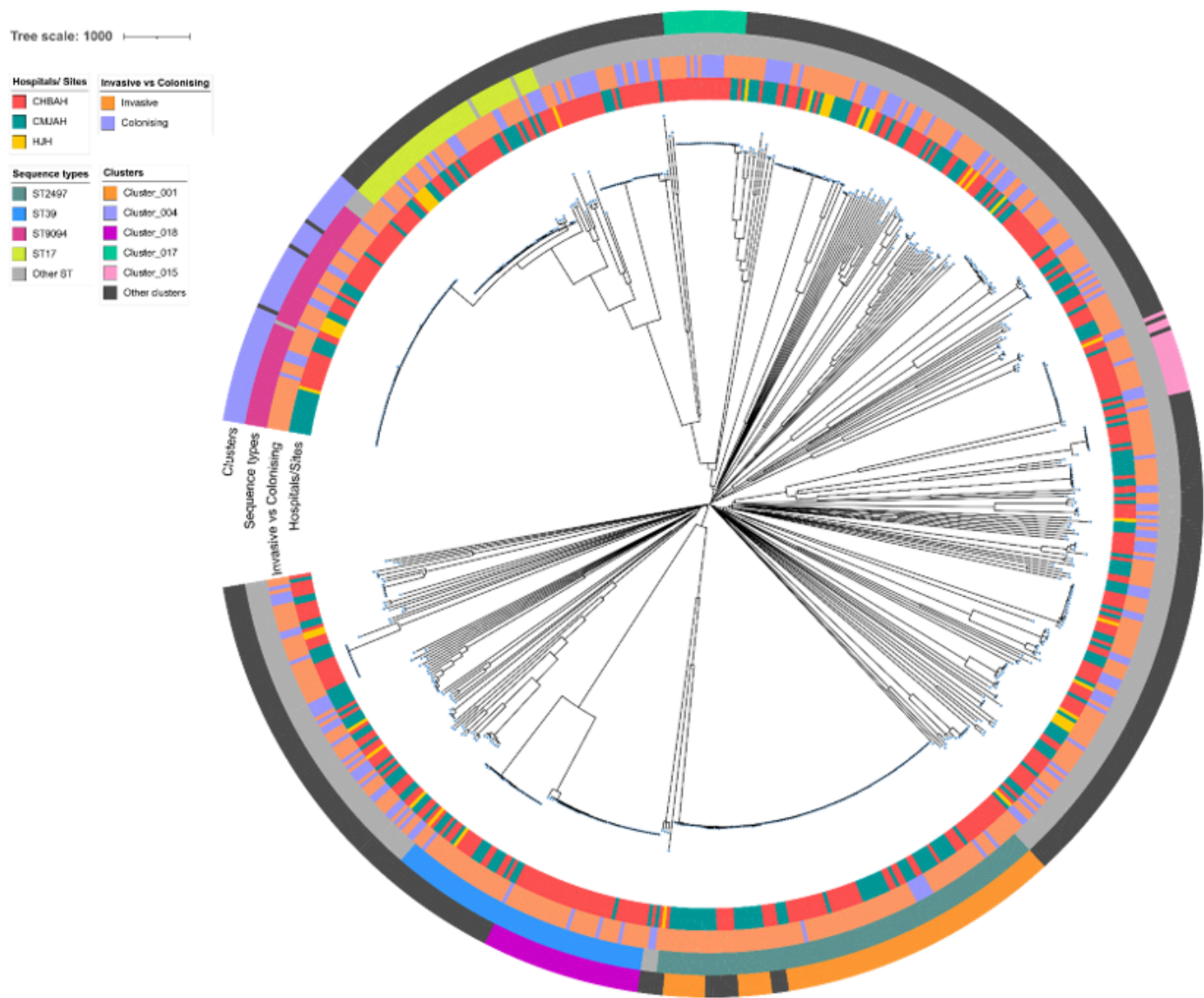


Figure 2

Midpoint rooted phylogenetic tree of *K. pneumoniae* invasive and colonising isolates in hospitalised adults in Johannesburg, South Africa (N=692, 2023-2024). Reference-free, distance-based phylogenetic trees were generated from de novo genome assemblies using the skatree function implemented in Split Kmer Analysis (SKA v2.0). The resulting Newick-formatted tree was visualised and annotated using iTOL (v7.0). The rings from inside out indicate (i.) Hospitals/ Sites, (ii.) Invasive versus Colonising isolates (iii.) Sequence types (ST), and (iv.) Clusters. The scale bar indicates genetic distance derived from split k-mer comparisons.

The hospital sites were Chris Hani Baragwanath Academic Hospital (CHBAH), Charlotte Maxeke Johannesburg Academic Hospital (CMJAH), and Helen Joseph Hospital (HJH). The STs depicted were

the four most common identified in invasive and colonising isolates: ST2497, ST39, ST9094, and ST17. The clusters depicted were the five largest arbitrarily named clusters identified: Cluster_001, Cluster_004, Cluster_018, Cluster_017, and Cluster_015.

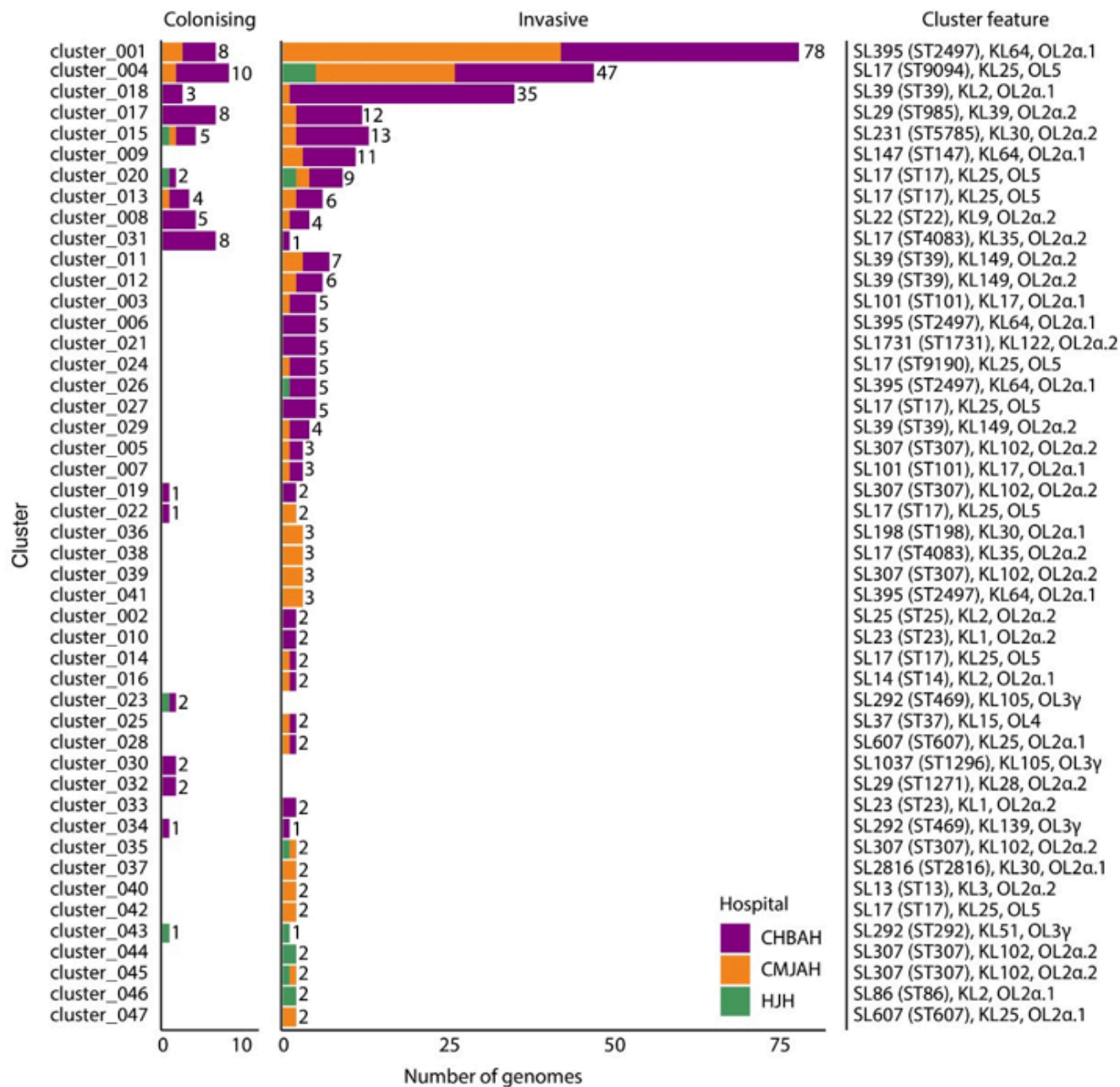
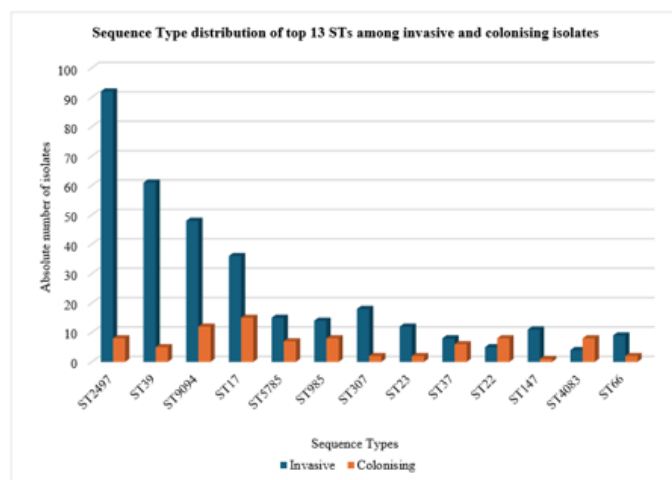


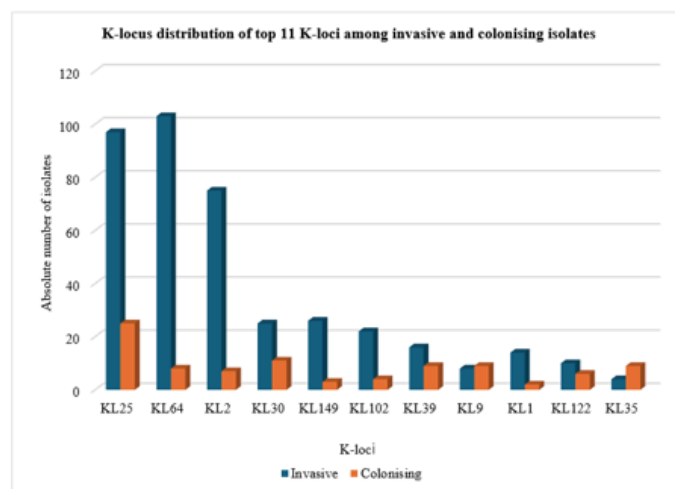
Figure 3

Distribution of genomically defined transmission clusters across the 692 isolates. Bar plot showing sizes of the 47 clusters (n=380 isolates, 54.9%) identified in invasive (n=317) and colonising (n=63) isolates defined as genomes differing by <10 PathogenWatch SNPs. Bars are coloured by hospital and annotated with sublineage (SL), sequence type (ST), and K and O loci.

4a



4b



4c

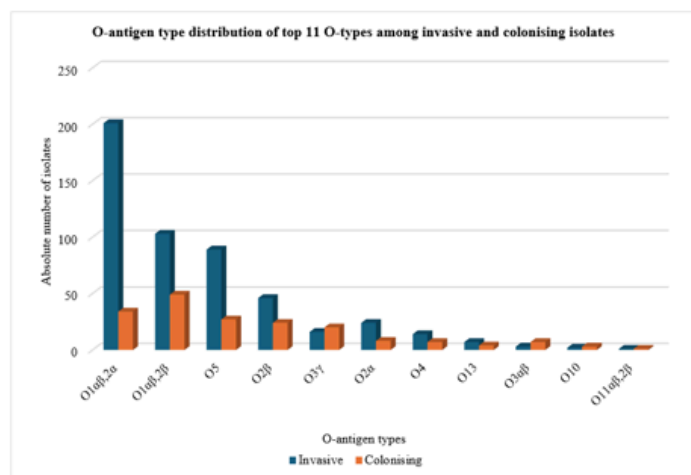


Figure 4

Distribution of genomic variables in invasive (n= 508) isolates and colonising (n= 184) isolates of *K. pneumoniae* that underwent genomic analysis from hospitalised South African adults. (a) Sequence type. (b) K-loci. (c) O-antigen type

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementarymaterialinvasivenessindexKPIDAug2026.xlsx](#)
- [Tables.docx](#)