

Genomic signatures and evolutionary dynamics of antimicrobial resistance revealed from 3,163 *Klebsiella pneumoniae* genomes

Hongru Li

First Affiliated Hospital of Anhui Medical University

Sifei Wang

Xi'an Jiaotong University

Jiabin Li

First Affiliated Hospital of Anhui Medical University

Han Xia

Hugobiotech Co., Ltd.

Kai Ye

Xi'an Jiaotong University

Yanyan Liu

First Affiliated Hospital of Anhui Medical University

Xiaofei Yang

xfyang@xjtu.edu.cn

Xi'an Jiaotong University

Research Article

Keywords:

Posted Date: September 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10893567/v1>

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

[Read Full License](#)

Additional Declarations: Competing interest reported. The findings described in this manuscript are the subject of a Chinese patent application. The authors declare no other competing interests.

Abstract

Background

Antimicrobial resistance (AMR) in *Klebsiella pneumoniae* is a critical global health threat, yet the genomic signatures and evolutionary dynamics underlying resistance—beyond known resistance genes—remain incompletely characterized. A comprehensive understanding of how both core-genome variation and accessory genomic elements collectively shape resistance phenotypes is urgently needed to inform surveillance and therapeutic strategies.

Methods

We generated a genomic dataset of 1,665 clinical *K. pneumoniae* isolates collected through a 15-year longitudinal surveillance network (HuiNet, 2006–2020), each with matched antimicrobial susceptibility profiles for 17 clinically relevant antibiotics across eight drug classes. Integrated with 1,498 public genomes, this resource yielded a pan-genome of 3,163 genomes. We systematically captured dispensable gene presence/absence and non-synonymous core-genome variants, and applied an ensemble machine-learning framework to identify features associated with resistance across all 17 antibiotics. Structural consequences of prioritized variants were evaluated using AlphaFold3 and Missense3D, and temporal evolutionary dynamics were examined across the collection period with independent geographic validation.

Results

The HuiNet cohort exhibited extensive resistance diversity, with 74.9% of isolates classified as multidrug-resistant. Pan-genome analysis revealed 3,712 core genes and 18,866 dispensable genes, with known AMR genes distributed across both compartments. We identified 369 AMR-associated features significantly enriched in resistance-related pathways. Based on these features, we achieved high predictive performance across all 17 antibiotics (mean AUC = 0.92), substantially outperforming models using only known determinants. Structural analyses revealed specific protein alterations associated with altered susceptibility, including missense mutations and coding repeat variations. Critically, an *ISEcp1-bla_{CTX-M-15}* transposition unit progressively replaced *bla_{CTX-M-14}*-associated genomic configurations over the 15-year period, with its association with cephalosporin resistance strengthening over time—a trend independently validated in a geographically distinct cohort.

Conclusions

This study establishes a scalable, pan-genome-based framework for linking bacterial genotype to AMR phenotype and demonstrates that resistance evolution shapes the resistance landscape through both

mobile-element configurations and coding variation in core genes, not merely through the acquisition of individual resistance determinants. These findings have implications for genomic surveillance and for understanding the evolutionary dynamics that shape clinical resistance landscapes.

Background

Klebsiella pneumoniae (*K. pneumoniae*) is a major cause of antimicrobial-resistant healthcare-associated infections, neonatal sepsis¹, and liver abscess^{2,3}. Whole-genome surveillance reveals extensive heterogeneity in clonal backgrounds, resistance phenotypes, and mobile genetic elements⁴, complicating genotype-phenotype mapping^{5,6}. While acquired β -lactamase genes are the major resistance determinants for carbapenemases⁴, isolates carrying identical genes often exhibit divergent resistance levels, indicating that gene presence alone cannot fully explain phenotypic variation^{3,7,8}. Chromosomal mutations affecting porins^{8–10}, envelope permeability^{10,11}, and efflux¹⁰ further modulate resistance. Thus, antimicrobial resistance (AMR) in *K. pneumoniae* is best framed as a pan-genome problem involving both accessory genes and core-genome variants^{3,6,12}.

However, a comprehensive view of AMR-associated variation remains elusive^{13–16}. Studies designed around predefined resistance gene panels¹³ or selected high-risk clones^{17,18} have been instrumental in characterizing the major β -lactamases and clonal expansions that dominate the AMR landscape. Yet their focused design can underrepresent genomic variation dispersed across the broader population, particularly core-gene variants and accessory elements in non-dominant lineages that may collectively contribute to resistance phenotypes. This underrepresentation is compounded by a tendency to reduce mobile-element-embedded resistance loci to isolated gene determinants, overlooking the local gene organization that can influence resistance expression or dissemination^{15,19,20}. Furthermore, genome-wide analyses often treat genes, variants, and mobile-element contexts as separate signals rather than as interconnected components of AMR architecture^{13,14}, fragmenting an inherently unified problem into disconnected feature types. Additionally, even when novel features are discovered, their structural consequences on encoded proteins²¹ and their long-term evolutionary dynamics are rarely explored, leaving a substantial gap between statistical prediction and mechanistic understanding.

To address these gaps, we generated genomic dataset of 1,665 clinical *K. pneumoniae* isolates collected through a 15-year longitudinal surveillance network (HuiNet, 2006–2020), with matched antimicrobial susceptibility profiles for 17 clinically relevant antibiotics across eight drug classes. Integrated with 1,498 publicly available genomes from PATRIC²², this resource yielded a pan-genome of 3,163 genomes, from which we systematically captured dispensable genes and non-synonymous core-genome variants. A two-stage machine-learning model identified 369 AMR-associated features across all 17 antibiotics, achieving high predictive performance with mean AUC of 0.92 and substantially outperforming models using only known determinants. Mechanistic characterization of prioritized variants revealed specific protein alterations associated with altered susceptibility, including missense mutations and coding repeat variations. Cephalosporin resistance was linked to distinct gene configurations centered on

ISEcp1 and *bla*_{CTX-M} variants. Notably, an ISEcp1-*bla*_{CTX-M-15} transposition unit progressively replaced *bla*_{CTX-M-14}-associated patterns over the 15-year period, and this replacement was associated with a sustained increase in cephalosporin resistance that was recapitulated in an independent geographic cohort.

Methods

Bacterial isolates and antimicrobial susceptibilities

Bacterial isolates were obtained from the multicenter hospitals in Anhui Province, China. The total of 1964 *K. pneumoniae* isolates were collected from routine microbiological cultures of clinical samples (urine, blood, sputum, tissue biopsies, etc) during 2006–2020. The antimicrobial susceptibility profiles of the 1,665 isolates were determined for 17 commonly used clinical antibiotics across different classes. These included: aminoglycosides (amikacin (AMK); gentamicin (GEN); tobramycin (TOB)), cephalosporins (ceftazidime (CAZ); cefuroxime (CXM); cefazolin (CZO); cefoxitin (FOX)), fluoroquinolones (ciprofloxacin (CIP); levofloxacin (LVX)), carbapenems (ertapenem (ETP); imipenem (IPM); meropenem (MEM)), penicillin combinations (ampicillin-sulbactam (SAM); piperacillin-tazobactam (TZP)), folate pathway inhibitors (trimethoprim-sulfamethoxazole (SXT)), tetracyclines (tetracycline (TCY)), and nitrofurans (nitrofurantoin (NIT)). Antibiotic abbreviations and clinical breakpoints were adopted from the Clinical and Laboratory Standards Institute (CLSI) M100 30th Edition, and minimum inhibitory concentrations (MICs) were then converted to susceptible (S), resistant (R) or intermediate (I) determinations. Multidrug resistance (MDR) was defined as non-susceptibility to three or more classes of antimicrobials. Extensively drug-resistant (XDR) was defined as non-susceptibility to at least one agent in all but two or fewer antimicrobial categories²².

Illumina sequencing

The genomic DNA of the clinical isolates was extracted using the SDS method with a QIAamp DNA Mini Kit (Qiagen) and quantified by a Qubit® 2.0 Fluorometer (Thermo Fisher Scientific, USA). The DNA was fragmented to a size of 350 bp by sonication, and sequencing libraries were prepared using the NEBNext® Ultra™ DNA Library Prep Kit. Paired-end sequencing with 150-bp mode was conducted on the Illumina NovaSeq 6000 platform (Illumina Inc., USA), with an average sequencing depth of approximately 122×.

Nanopore sequencing

Genomic DNA from a subset of clinical isolates was extracted for Nanopore sequencing using the ZymoBIOMICS DNA Miniprep Kit D4300 following the manufacturer's recommendations. Sequencing libraries were prepared using Oxford Nanopore Technologies's 1D ligation sequencing kit (SQK-LSK109) with native barcoding (EXP-NBD196) according to the manufacturer's instructions. The sequencing

libraries were loaded onto R9.4.1 flow cells and sequenced using a GridION X5 platform, with an average sequencing depth of approximately $60\times^{23}$.

Genome assembly and assessment

For isolates sequenced using only Illumina sequencing, *de novo* assembly of the clean reads was performed with Unicycler v0.4.8²⁴ using the following parameters: `--min_fasta_length 500, -t 20, --keep 2`. For isolates that underwent both Illumina and ONT sequencing simultaneously, a strategy involving assembly by ONT sequencing data followed by polishing with Illumina data was used. Several software programs, including Unicycler v0.4.8 (`--min_fasta_length 500, -t 20, --keep 2`), Canu v2.1.1²⁵ (`genomeSize=4m, corOutCoverage = 40`), Flye v2.8.3²⁶ (`--nano-corr, --genome-size 4m, --threads 8, --iterations 1`), and wtdbg2 v2.5²⁷ (`-t 8, -x corrected, -g 4m`), were used for assembly of the Nanopore long reads separately. Then, the draft genomes with the highest N50 values were polished with corrected long reads using Racon v1.4.20²⁸ (`-u, -t 8`) and high-quality Illumina paired-end short reads using Pilon v1.23²⁹ (`--changes, --fix all`) for three iterations. The polished sequences were then circularized using Circlator v1.5.5³⁰ with default parameters. The genome quality were assessed using BUSCO v5.2.1²³ (`-m genome`) against bacterial_odb10 library dataset. Only isolates identified as *K. pneumoniae* by Kleborate v0.4.0-b³ were retained for epidemiological and statistical analyses.

Genotyping, AMR and virulence prediction

The MLST tool was applied to determine sequence types (STs) using the *K. pneumoniae* MLST scheme³¹ available via the Pasteur database. Abricate v1.1.2 (<https://github.com/tseemann/abricate>) was used to detect acquired antimicrobial resistance genes (ARGs) using the ResFinder³² database. The associations between antimicrobial resistance genes and antibiotics were obtained from the ResFinder database. To analyze temporal trends of ARGs across different antibiotic categories, we proceeded as follows: (i) compile a list of ARGs associated with each antibiotic category based on the resistance gene-antibiotic relationships provided by Resfinder; (ii) count the number of isolates carrying these ARGs for each year; and (iii) calculate the relative proportion by dividing the number of isolates carrying each ARG by the total number of isolates in that year.

The plasmid replicons were predicted by PlasmidFinder³³. Completed plasmid sequences were compared pairwise using two similarity metrics: Mash similarity was calculated with Mash³⁴ and gene content Jaccard similarity was determined using Roary³⁵. Plasmids pairs with Mash similarity ≥ 0.98 and Jaccard similarity ≥ 0.8 were designated as conserved plasmids. The genetic background of acquired AMR and virulence genes was investigated with Bandage v0.8.1³⁶. Kleborate v0.4.0-b was used for the capsular polysaccharide (K) and lipopolysaccharide O-antigen serotype prediction based on *wzi* alleles and Kaptive, detection of *ICEKp* associated virulence loci (*yersiniabactin* [*ybt*], *colibactin* [*clb*]), virulence

plasmid associated loci (salmochelins [iro], aerobactin [iuc], hypermucoidy [rmpA, rmpA2]), and assessment of the *ompK35/ompK36* gene integrity.

Core-genome SNP-based phylogenetic reconstruction for epidemiological characterization

For epidemiological characterization of the HuiNet isolates, coding sequences were predicted using Prodigal v2.6³⁷ and annotated with Prokka v1.13.3³⁸. Core genes were identified using Roary v3.12³⁵, and a concatenated core-gene alignment was generated with the options "--e --mafft". Core-genome single-nucleotide polymorphisms (cgSNPs) were extracted from the alignment using SNP-sites³⁹. Isolate pairs differing by fewer than four cgSNPs were considered clonal. Recombinant regions were identified and removed using Gubbins v2.3.4⁴⁰. A maximum-likelihood phylogeny was then inferred from the recombination-filtered alignment using RAxML v8.2.10⁴¹ under the GTRGAMMA model. The resulting tree was visualized and annotated using iTOL⁴². This Roary-based cgSNP workflow was used solely to characterize population structure in the epidemiological analysis and was independent of the variant-feature extraction pipeline used in downstream AMR modeling.

Statistical analysis of epidemiological data

Statistical analyses for the epidemiological characterization were performed in R v4.1.0. Categorical variables were compared using Chi-square tests or Fisher's exact tests, as appropriate. Correlations were assessed using Pearson correlation coefficients. Unless otherwise stated, a two-sided $P < 0.05$ was considered statistically significant.

Construction of the pangenome

The pangenome of *K. pneumoniae* was constructed using Panaroo v1.5.2⁴³ (--clean-mode strict, --alignment pan, --aligner mafft, --core_threshold 0.9, -t 48, --refind_prop_match 0.98, --search_radius 1000). Annotated genomes were clustered based on sequence homology, with each cluster representing a gene. Genes present in more than 90% of isolates were defined as core genes, whereas the remaining genes were considered dispensable. To ensure cluster quality, sequences within each gene cluster were filtered as follows:

1. The mode of sequence lengths for each cluster was calculated, and sequences deviating by more than ± 100 bp or 20% of the mode were removed.
2. Clusters whose reference sequence was removed in step 1 were entirely excluded from downstream analyses.

Core-gene variant identification and landscape analysis for AMR modeling

High-confidence core genes were subjected to single-base-resolution variant calling. Reads were aligned to core gene references using BWA v0.7.17-r1188⁴⁴, followed by variant calling, sample merging and multiallelic variants record splitting with BCFtools v1.15.1⁴⁵. For each gene, graph-based pangenomes were constructed using vg v1.60.0⁴⁶ based on the combined alignments across all samples, and the resulting graphs were visualized. Variants were classified as synonymous, missense, or other types based on the annotation result of SnpEff v5.2⁴⁷.

Functional enrichment analyses

To explore the potential biological relevance of core-gene variation at the landscape level, genes harboring core-gene variants predicted by SnpEff to have high-impact were subjected to KEGG⁴⁸ enrichment analysis. Separately, to interpret the functional composition of the AMR candidate signals identified by the downstream framework, KEGG enrichment analysis was also performed on the full candidate feature set, including all dispensable genes corresponding to candidate g-features, all genes harboring candidate core-gene features.

Pan-genomic feature matrix extraction

The presence/absence of g-features and v-features together formed the pan-genomic feature matrix. Presence/absence calls for v-features were derived from VCF files, whereas those for g-features were obtained directly from the Panaroo output. Features were encoded as 1 for presence and 0 for absence. To retain biological significance, features were filtered as follows:

1. Features present in fewer than four isolates were removed.
2. Synonymous substitutions in core genes were excluded.

For each antibiotic, isolates exhibiting R phenotype were labeled as 1, those with S phenotype as 0, and isolates with I phenotype were excluded.

Data partitioning

The full dataset ($n = 3,163$ genomes) was divided into training, testing, and validation sets. A combined pool consisting of all 1,665 HuiNet isolates and 50% of the PATRIC isolates was randomly shuffled and split into a 70% training set and a 30% testing set. The remaining 50% of the PATRIC isolates were reserved as an independent validation set to assess model generalizability.

Significant AMR pan-genomic feature selection

To identify pan-genomic features significantly associated with antibiotic resistance, we used a two-step selection strategy. First, the association between each feature and the binary AMR phenotype was

evaluated using both Chi-square and Kruskal-Wallis tests, and only features with $p < 0.05$ in both tests were retained.

Second, for each antibiotic i , the retained features were ranked independently by four machine learning models—logistic regression (LR), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB)—implemented using scikit-learn v1.6.1⁴⁹ and XGBoost v2.1.4⁵⁰. Ranking was based on the absolute values of model coefficients for LR and SVM, and on feature importance scores for RF and XGB.

The optimal ranking threshold, K_i , was determined within an expanding search interval $[1, K_{med}]$. Specifically, K_{med} was increased iteratively with a step size of 1, and the area under the receiver operating characteristic curve (AUC) on the testing set was monitored to identify the local optimum. This procedure was designed to balance model parsimony, targeting a refined feature set of approximately 20–50 features, with predictive performance. A feature was ultimately defined as a significant AMR-associated pan-genomic feature if it ranked within the optimized K_i in at least two of the four models.

Ensemble model construction and performance validation

Using the significant features selected above, we then trained a soft-voting ensemble composed of the same four base classifiers: LR, RF, SVM, and XGB. Hyperparameters were optimized by randomized search with five-fold cross-validation on the training set only. The search space included C for LR and SVM, $n_estimators$ and max_depth for RF and XGB, $learning_rate$ for XGB, and the integration weights for the soft-voting ensemble. After fixing the optimal hyperparameters and the feature threshold K_i , model performance was assessed on the testing set and further evaluated on the independent validation set using AUC.

Identification of known AMR gene features and classification of all representative features

Based on the annotations obtained by integrating results from the CARD⁵¹, ResFinder³², and NCBI⁵² databases using Abricate v1.1.2, known AMR-related dispensable genes' cluster names were mapped with their official names. Known AMR-related single-nucleotide polymorphisms (SNPs) were annotated using RGI v6.0.4 (main, -t contig, --clean). To determine the function of the gene corresponding to each feature, the central sequence of the corresponding cluster was manually aligned against the NCBI database using BLAST⁵³, and the best hit was selected based on the smallest possible E-value, the closest match in sequence length, and the continuity of the alignment to ensure accurate functional inference of the corresponding protein.

Performance evaluation using feature subset ablation

To evaluate the contribution of different categories of features, we constructed three subset feature matrices for each antibiotic-specific model: (i) v-features only, (ii) g-features only, and (iii) previously known AMR features only. Models were retrained on each feature subset as well as on the full feature

set, following the same procedure as described below: hyperparameters were optimized on the training dataset using five-fold cross-validation combined with randomized search. The ensemble model was then refitted on the entire training set using the selected hyperparameters. The refitted model was applied to the internal test dataset and to the external PATRIC validation dataset to generate predicted probabilities, from which AUC values were calculated.

To assess whether the full feature set provided a performance advantage over the subset feature sets, AUC values across the 17 antibiotics were compared between feature categories. For each evaluation dataset (test and validation) and each feature category, the AUCs from the 17 antibiotic-specific models were treated as one group. Differences between the full feature set and each subset were assessed using the Mann–Whitney U test. Since the NIT’s model contained no previously known AMR features, comparisons involving the “known-features” subset were based on 16 antibiotics instead of 17.

Permutation-based feature ranking within antibiotic-specific feature sets

For each antibiotic, feature ranking was performed within its corresponding full feature set, including both v-features and g-features, using permutation importance⁵⁴. The previously trained soft-voting ensemble model was fixed and applied without retraining. Importance was computed on test dataset and verified on the validation dataset. For each feature, its values were randomly permuted across samples while all other features were kept unchanged, and the resulting change in predictive performance was recorded. This procedure was repeated multiple times per feature to obtain stable estimates. The mean decrease in performance across repetitions was used as the importance score, and features were ranked in descending order according to the absolute mean importance values. Permutation importance was implemented in Python using the `permutation_importance` function from `scikit-learn` v1.7.2⁴⁹.

Feature-based sample clustering and marker identification

Samples were clustered using the `pheatmap` function from the R `pheatmap` package with “`cluster_methods = ward.D2`” parameter, based on their similarity patterns across prioritized features. Marker features were manually selected by visually inspecting the resulting heatmaps. Log-odds ratios (LORs) were computed directly from the feature and phenotype matrices in Python. A continuity correction was applied to the 2×2 contingency tables by adding 0.5 to each cell, in order to mitigate the impact of zero counts and stabilize the estimates.

Structural analysis of prioritized core-gene single-nucleotide variants

To investigate the potential structural consequences of the identified core-gene single-nucleotide variants (SNVs), we performed a combined in silico analysis using AlphaFold3⁵⁵ (<https://alphafoldserver.com/>) and Missense3D⁵⁶ (<https://missense3d.bc.ic.ac.uk>) for further analysis of

top-ranked previously unidentified variants. For each antibiotic, features were ranked by permutation importance, and subsequent analysis focused on the top 15 ranked v-features, particularly those not previously documented as established AMR markers.

The structural impact of these novel SNVs was assessed through a multi-step pipeline. First, the protein sequences of the unique representative gene cluster centroids (longCentroid in Panaroo) containing the target variants were modeled using AlphaFold3 to obtain high-resolution three-dimensional structures. To ensure the reliability of downstream analysis, we filtered variants according to the predicted local distance difference test (pLDDT), retaining only mutations located in regions with high confidence in the predicted local structure (pLDDT ≥ 70). Subsequently, these high-pLDDT structures were submitted to the Missense3D online server to evaluate the stereochemical consequences of the specific amino acid substitutions. Protein domain annotation was performed using InterPro (<https://www.ebi.ac.uk/interpro/>) via its web interface.

Contig-level positional pattern analysis of selected genes

For the five selected genes, *ISEcp1*, *group_2412*, *bla*_{CTX-M-15}, *group_12673*, and *bla*_{CTX-M-14}, contig-level positional patterns were defined in each isolate based on their genomic order within the same contig. Genes located on different contigs were not combined into a single pattern. Within each contig, genes were ordered by genomic coordinates, and positional patterns were defined as contiguous blocks of one or more target genes; multiple target genes were assigned to the same pattern only when they were immediately consecutive, with no intervening non-target genes. Reverse orders were treated as equivalent and collapsed into one canonical pattern. The above analyses were performed in Python, and visualization was performed in R using gggenes⁵⁷ v0.6.0 as a ggplot2⁵⁸ extension.

Temporal analysis of positional pattern prevalence and resistance association

For temporal analyses, isolates were grouped by collection year. For each positional pattern, the yearly prevalence was calculated as the number of isolates carrying the pattern divided by the total number of isolates collected in the corresponding year. Temporal changes in pattern prevalence were assessed using the Cochran–Armitage Test for trend⁵⁹, with collection years treated as ordered categories, and two-sided *p*-values were used to evaluate statistical significance. Prevalence analyses and visualization were performed in R, while the trend test was performed in Python.

To evaluate associations between positional patterns and resistance phenotypes, each pattern was encoded at the isolate level according to whether it was observed on at least one contig in that isolate. Associations between pattern presence and binary resistance phenotypes were evaluated using the LOR calculation described above. To further assess whether these associations changed over time, year-specific 2×2 contingency tables were constructed for each pattern–antibiotic pair. In each year, isolates were stratified by pattern presence or absence and by binary resistance phenotype. Yearly LORs were

then calculated from these contingency tables. For each pattern–antibiotic pair, Pearson correlation coefficients were calculated between yearly LOR values and collection year to summarize the direction and strength of temporal changes in resistance association, with corresponding *p*-values reported for the correlation tests. The above analyses and visualization were performed in R.

Results

Characterization of 1,665 *K. pneumoniae* isolates from HuiNet

We analyzed 1,665 *K. pneumoniae* isolates and their AMR phenotypes for 17 antibiotics collected through the HuiNet from 2006 to 2020. All 1,665 isolates underwent high-depth whole-genome sequencing using Illumina short reads, and 491 isolates additionally underwent Oxford Nanopore long-read sequencing for hybrid assembly (mean coverage, 348×). This produced 1,174 short-read assemblies and 491 hybrid assemblies, with the two assembly strategies unevenly distributed across the study period: All isolates collected from 2006 to 2016 were sequenced using Illumina alone, whereas all isolates collected from 2017 to 2019 were sequenced using the hybrid approach; among isolates collected in 2020, 175 were sequenced using Illumina alone and 62 were hybrid sequenced (**Supplementary Table 1**). Short-read assemblies achieved contig N50 values ranging from 0.01 Mb to 1.00 Mb, with a mean of 0.31 Mb, and recovered a median of 5,085 annotated genes per genome (range, 4,638–6,295). By comparison, hybrid assemblies yielded contig N50 values ranging from 0.18 Mb to 5.61 Mb, with a mean of 3.81 Mb, and recovered a median of 5,179 annotated genes per genome (range, 4,650–6,959) (**Supplementary Fig. 1, Supplementary Table 1**).

Antimicrobial susceptibility testing across eight antibiotic classes revealed diverse and frequently overlapping resistance profiles (**Supplementary Table 2**). Based on resistance to three or more antibiotic classes, 1,247 isolates (74.89%) were classified as MDR. Carbapenems retained the greatest clinical activity, with 91.17% of isolates susceptible to MEM. In contrast, high resistance rates were observed for β -lactam/ β -lactamase inhibitor combinations and cephalosporins, notably SAM (72.79% resistant), CXM (67.75% resistant), and CZO (63.82% resistant). Resistance also exceeded 60% for NIT (69.43%), TCY (64.44%), CIP (64.38%), and SXT (60.66%), underscoring the substantial MDR burden in this collection (**Supplementary Fig. 2, Supplementary Table 3**). Together, the high prevalence of MDR and the diversity of resistance profiles provided a broad phenotypic basis for the subsequent analysis of AMR-associated genomic features.

The longitudinal structure, genomic diversity, and heterogeneous resistance phenotypes of the HuiNet collection provided the basis for subsequent analyses (Fig. 1). cgSNP phylogeny revealed a genetically diverse population comprising many STs. ST15 (7.63%), ST23 (7.39%), ST37 (6.55%), and ST11 (6.13%) predominated, together accounting for 27.69% of isolates; an additional 16 STs, including ST412, ST65, ST29, ST35, and ST307, each comprised more than 15 isolates and together contributed a further 24.14% (**Supplementary Table 4**). ST distribution shifted significantly over the 15-year study period ($p = 3.20 \times 10^{-7}$, KW-test) (**Supplementary Fig. 3A**). Consistent with this lineage diversity, K-locus (KL) and O-

antigen typing also revealed considerable diversity and temporal variation. Among the 73 KL types identified, KL1 (8.47%), KL2 (5.95%), KL57 (4.44%), KL64 (3.54%), and KL102 (3.00%) were the most frequent (**Supplementary Table 4**). O serotyping identified O1v2 (26.37%), O1v1 (18.92%), O2v2 (12.73%), O3b (11.77%), and O2v1 (7.87%) as the dominant types (**Supplementary Table 4**). Both KL and O distributions changed significantly over time ($p = 2.03 \times 10^{-10}$ and $p = 2.62 \times 10^{-11}$, respectively, KW-test) (**Supplementary Fig. 3B,C**) and exhibited strong ST-specific associations, e.g., ST23 with KL1 and O1v2, ST11 with KL64 and O2v1 (**Supplementary Figs. 4, 5; Supplementary Table 4**).

Pan-genome of 3,163 *K. pneumoniae* genomes

To comprehensively study the genomic characteristics of *K. pneumoniae*, we combined our 1,665 genomes with 1,498 publicly available PATRIC genomes, yielding a pan-genome of 3,163 *K. pneumoniae* genomes for analysis. The pan-genome comprised 3,712 core genes (present in > 90% of genomes), 18,866 dispensable genes (present in > 3 and < 90% of genomes), and 6,612 rare genes (present in ≤ 3 genomes) (Fig. 2A). Among these, 268 known AMR genes were identified, including 22 in the core genes and 187 in the dispensable genes (Fig. 2B), confirming that resistance determinants span both genomic compartments (**Supplementary Table 5**). We catalogued variation within core genes and identified 655,351 SNPs, 1,130 insertions, and 1,694 deletions. Of these, 234,930 SNPs, 192 insertions, and 215 deletions were shared between the HuiNet and PATRIC collections. The HuiNet genomes harbored an additional 256,998 SNPs, 728 insertions, and 1,184 deletions as dataset-specific variants, while the PATRIC genomes contributed 163,423 SNPs, 210 insertions, and 295 deletions (Fig. 2C). Functional annotation revealed that synonymous substitutions accounted for 66.6% of all variants, missense mutations for 32.1%, and other variant types for 1.3% (Fig. 2D). Among the most severe coding changes, we identified 520 frameshifts, 115 start-loss events, and 1,138 stop-gain events. Genes carrying these high-impact variants were significantly enriched in pathways linked to AMR (e.g., β -lactam resistance, cationic antimicrobial peptide resistance), biofilm formation, quorum sensing, two-component systems, and central metabolism (Fig. 2E-G). Rather than being randomly distributed, high-impact core-gene variants were thus concentrated in functions that govern drug susceptibility, surface-associated growth, and environmental sensing. This observation suggests that core-gene variation provides a reservoir of functionally relevant diversity that may shape resistance phenotypes, motivating our subsequent systematic identification of AMR-associated features through machine learning.

AMR genomic features for 17 antibiotics

We sought to identify the genomic features that underpin resistance to each of the 17 antibiotics, using the pan-genome as a comprehensive source. Dispensable gene presence/absence was encoded as g-features, and non-synonymous core-gene variants (SNPs, insertions, and deletions) as v-features, thus capturing both the accessory genome and functional variation in the conserved genomic backbone. After removing low-frequency features, 18,866 g-features and 71,534 v-features were retained.

To distil a predictive and interpretable feature set, we applied a two-stage prioritization pipeline (Fig. 3A). First, features with statistically significant associations with resistance phenotypes ($p < 0.05$ for both Chi-square and Kruskal-Wallis tests) were retained, yielding 8,156 g-features and 38,732 v-features (**Supplementary Tables 6–9**). Second, we ranked these candidates by their importance across multiple machine-learning models, including LR, RF, XGB, and SVM, and selected those that appeared among the top-ranked features in at least two models (**Supplementary Figs. 6, 7**). This process produced a final panel of 369 AMR-associated features, including 242 g-features and 127 v-features (**Supplementary Table 10**). Forty-six of these features (6 v-features and 40 g-features) matched known resistance determinants recorded in CARD⁵¹, NCBI⁵², or ResFinder³², including the SHV G238S substitution used by NCBI Kleborate³ for AMR profiling (Fig. 3B; **Supplementary Tables 11–13**). Beyond these database-confirmed features, an additional subset was supported by genomic-context evidence or involvement in AMR-related pathways. For example, one g-feature annotated as *klcA*_3, shared high similarity with the antirestriction gene *klcA*_{HS}, which resides in the *bla*_{KPC-2} genetic context of the China-associated ST11 plasmid lineage^{60–62}. Three genes harboring selected v-features and 16 selected g-features participate in β -lactam or cationic antimicrobial peptide resistance pathways. These features, although not yet catalogued in public AMR databases, are not isolated statistical associations but reside in genomic or functional contexts already linked to resistance, supporting their designation as putative AMR-associated features.

For each antibiotic, we constructed a soft-voting ensemble model combining LR, RF, SVM and XGB, trained and internally tested on the combined HuiNet and PATRIC data, and independently validated on the PATRIC collection (Fig. 3A). All 17 antibiotic-specific models achieved AUCs above 0.8 on both the internal test set and the external validation set (Fig. 3C; **Supplementary Fig. 8; Supplementary Table 14**). We then assessed the contribution of different feature types. Full-feature models achieved a mean AUC of 0.92 on the test set, significantly outperforming models using only v-features (mean AUC = 0.69; $p = 9.32 \times 10^{-7}$, one-sided Mann-Whitney U test) or only g-features (mean AUC = 0.82; $p = 2.54 \times 10^{-5}$). The similar pattern held on the independent validation set (full-feature mean AUC = 0.93 vs. v-only = 0.84 and g-only = 0.90) (Fig. 3D; **Supplementary Table 15**). Moreover, models built on the selected features consistently outperformed those using only the previously known AMR features, with higher mean AUCs on both the test set (0.92 vs. 0.83; $p = 2.54 \times 10^{-5}$) and the validation set (0.93 vs. 0.91; Fig. 3E). These results demonstrate that core-gene variants and dispensable genes contribute complementary and additive predictive information, and that the pan-genome-wide search captures resistance-associated signals that extend substantially beyond current databases.

Mechanisms of AMR genomic features

To investigate potential mechanisms suggested by in silico analyses, we examined the functions of the 369 prioritized features. We performed KEGG pathway enrichment on the genes harboring v-features and the genes represented by g-features, and found that both feature classes are significantly enriched in β -lactam resistance ($q = 5.76 \times 10^{-3}$ and $q = 5.56 \times 10^{-37}$, respectively), a core pathway containing canonical AMR genes (Fig. 4A; **Supplementary Tables 16–17**). In addition, starch and sucrose

metabolism ($q = 4.92 \times 10^{-5}$ for v-features) and folate biosynthesis ($q = 1.41 \times 10^{-2}$ for g-features) were also enriched (Fig. 4A; **Supplementary Tables 16–17**), likely reflecting metabolic contexts that modulate resistance rather than direct mechanisms. These results indicate that the model-prioritized features are anchored in known resistance biology and in functionally coherent metabolic backgrounds, providing a rich substrate for mechanistic dissection.

We used a permutation-based framework to rank individual features by importance within the ensemble model (**Supplementary Table 18**). Of the top five features per antibiotic, 60.0% corresponded to known AMR determinants or related regulators, underscoring the concordance of data-driven ranking with established mechanisms. We identified an SXT-associated missense variant (S178R) in *group_258*, which encodes a lysoplasmalogenase-like protein with a YhhN domain (Fig. 4B; **Supplementary Table 19**). Missense3D predicted a cavity-altering effect ($\Delta = -86.8 \text{ \AA}^3$) within this domain, and Missense3D-TM further predicted that the substitution introduces a charged residue in a membrane-associated environment, suggesting potential perturbation of local transmembrane architecture (**Supplementary Table 20**). The affected residue lies in a region with high pLDDT scores (> 90) in the AlphaFold3-predicted reference protein structure. The variant was strongly associated with SXT sensitivity (LOR = -2.67; 69 sensitive vs. 27 resistant isolates), and all isolates carrying this variant belonged to ST307 (**Supplementary Table 21**). Beyond SNVs, insertions also showed potential AMR mechanisms. We found that an insertion in the top-5 features for NIT was a repeat expansion (GCTCCG, encoding one AP unit) in a microsatellite region in *ompA*, encoding the major outer-membrane protein OmpA, a known AMR related protein in *K. pneumoniae*⁶³ (Fig. 4C; **Supplementary Fig. 9**). The expansion resulted in a five-copy allele that occurred in 20 isolates across 15 STs (**Supplementary Table 22**) and maps to a flexible linker between the β -barrel and the periplasmic domain^{64,65}, and was strongly associated with NIT sensitivity (LOR = -3.89), indicating that copy-number variation at coding microsatellites may represent an underappreciated mechanism modulating antimicrobial susceptibility in *K. pneumoniae*. Together, these results demonstrate that prioritized v-features capture structurally interpretable coding variations, expanding the spectrum of resistance-associated variation beyond canonical gain-of-resistance mutations.

Clustering analysis revealed clear phenotype separation for several antibiotic classes, most prominently for cephalosporins (CAZ/CXM/CZO), fluoroquinolones (CIP/LVX), carbapenems (IPM/MEM), and aminoglycosides (AMK) (Fig. 4D; **Supplementary Figs. 10–12**). Across 54 feature-antibiotic pairs in these classes, 98.15% showed positive LORs, and 87.04% had absolute LORs > 2 (**Supplementary Table 23**). For cephalosporins, nine features, including two SNVs and seven genes, effectively discriminated resistant from susceptible isolates. Among the newly identified genes, *group_2412* displayed consistently large positive LORs (CAZ: 4.36; CXM: 6.51; CZO: 5.38) (**Supplementary Table 23**). Notably, four of these nine cephalosporin-associated features, including *bla*_{CTX-M-15}, *bla*_{CTX-M-14}, *group_12673*, and *group_2412*, frequently co-occurred with the transposable element *ISEcp1*. Systematic analysis of their genomic neighborhood across all 3,163 isolates identified 17 distinct organization patterns (**Supplementary Table 24**). The three most frequent patterns in the HuiNet cohort were

designated P1, P2, and P3. P1 and P2 represented *bla*_{CTX-M-14}-associated configurations, whereas P3 contained the complete *ISEcp1-group_12673-bla*_{CTX-M-15-group_2412} module^{66,67} (Fig. 4E). The temporal dynamics of these patterns revealed a striking evolutionary shift. Over the 15-year HuiNet collection period, P1 and P2 showed significant decreasing trends (Cochran-Armitage Trend test, $Z = -3.25$ and -4.65 ; $q = 4.92 \times 10^{-3}$ and 2.77×10^{-5} , respectively), while P3 exhibited a significant temporal increase ($Z = 5.54$, $q = 5.06 \times 10^{-7}$) (Fig. 4F; **Supplementary Fig. 13; Supplementary Tables 25–26**). This replacement of *bla*_{CTX-M-14}-dominant patterns by the *ISEcp1-bla*_{CTX-M-15} module was tightly coupled to resistance: the yearly LOR of P3 for CAZ, CXM, and CZO rose significantly with collection year (Pearson $r = 0.66$, 0.67 , and 0.53 ; $p \leq 0.05$) (Fig. 4G; **Supplementary Tables 27–28**). The same directional trend was observed in an independent Houston PATRIC cohort (2011–2015; CAZ: $r = 0.34$; CXM: $r = 0.38$; CZO: $r = 0.97$) (**Supplementary Fig. 14; Supplementary Tables 29–32**). All three patterns were broadly distributed across the phylogeny rather than being restricted to a particular ST (**Supplementary Fig. 15**). Collectively, these findings reveal a long-term evolutionary shift in the cephalosporin-resistance landscape, in which the *ISEcp1-bla*_{CTX-M-15} transposition unit progressively replaced *bla*_{CTX-M-14}-associated configurations and drove a sustained increase in resistance over 15 years.

Discussion

We established the HuiNet surveillance network and collected 1,665 clinical isolates over a 15-year period, each with matched genomic-phenotypic data, providing a continuous longitudinal sampling window that is difficult to obtain from publicly archived datasets. Integrating these genomes with 1,498 public PATRIC genomes, we constructed a unified pan-genome framework for investigating AMR in *K. pneumoniae* that jointly captures dispensable genes and core-genome variants⁶⁸ across 3,163 genomes. We return to the central question: can a comprehensive pan-genome representation uncover resistance-associated features beyond known determinants, and can these features be mechanistically interpreted? Our findings provide an affirmative answer.

Models built on the full feature set consistently outperformed those using either feature class alone or only known AMR determinants, demonstrating that accessory genes and core-genome variants provide complementary information. Nearly over half of the top-ranked features per antibiotic matched established resistance determinants or related regulators, validating the biological relevance of our data-driven ranking. The remaining features, predominantly uncharacterized, substantially expand the searchable space of AMR-associated markers. We emphasize that these novel features should not be equated with direct causal determinants^{69,70}; some may reflect plasmid backgrounds^{71,72}, lineage-associated contexts⁷¹, or co-selected MDR states⁷³. Their primary value lies in providing a focused candidate set for downstream investigation, defining a searchable space that connects statistical association with testable mechanistic hypotheses^{69,70,74}.

The variant-level analyses further distinguish this study from purely predictive screens. The S178R substitution in *group_258* and the coding repeat expansion in *ompA* were both negatively associated with

resistance, demonstrating that prioritized features can mark susceptible-associated states rather than resistance-conferring alleles. The *ompA* finding is particularly noteworthy: it shows that our pipeline captures coding repeat-length variation in an outer-membrane protein with established AMR relevance, raising the possibility that tandem repeat variation represents an underappreciated mechanism modulating susceptibility in Gram-negative pathogens⁷⁵. Together, these examples demonstrate that prioritized v-features are not merely statistical correlates but can point to biologically meaningful variation with measurable structural and phenotypic consequences.

Beyond individual variants, our results reveal that resistance-associated genomic architecture in *K. pneumoniae* is dynamic and can undergo long-term directional shifts. The three predominant cephalosporin-associated gene organization patterns displayed a clear temporal replacement: P1 and P2 (*bla*_{CTX-M-14}-associated) declined significantly over 15 years, while P3—the *ISEcp1-group_12673-bla*_{CTX-M-15-group_2412} transposition unit—expanded significantly. This shift was tightly coupled to rising cephalosporin resistance, with P3's yearly LOR increasing significantly for CAZ, CXM, and CZO, and the same trend was recapitulated in an independent Houston cohort. We propose that this replacement represents a genuine evolutionary transition in which the genomic organization surrounding a key resistance gene, rather than the gene alone, constitutes the unit of selection, potentially reflecting the ability of *ISEcp1*-linked contexts to couple *bla*_{CTX-M} mobilization with cis-regulatory effects such as promoter provision^{19,20,66}. This finding extends the conventional focus on clonal expansion or horizontal gene transfer of individual resistance determinants⁷⁶, and suggests that surveillance programs should monitor not only AMR gene presence but also the mobile-element contexts that facilitate their dissemination⁷⁷.

Several limitations merit consideration. Our model prioritizes associations over causal inference, necessitating experimental validation for most novel features; we therefore emphasize that our prioritized features represent correlative signals requiring functional validation to establish definitive causality, with a subset supported by structural prediction or evolutionary trend analysis. The observed temporal replacement of the *ISEcp1-bla*_{CTX-M-15} module may be confounded by clonal background, although the same directional trend in an independent Houston cohort and its occurrence across multiple STs (**Supplementary Fig. 15**) mitigate this possibility. The HuiNet cohort is geographically restricted to Anhui Province, China, and global generalizability awaits further sampling, though our analytical pipeline is publicly available for analogous studies. Additionally, while *ISEcp1*-mediated transposition of *bla*_{CTX-M} is mechanistically well understood, our contribution lies in quantifying its long-term population-level replacement dynamics over 15 years, revealing a directional shift in the resistance landscape that underscores the importance of tracking mobile-element contexts—not just resistance genes—in cephalosporin resistance surveillance.

Conclusions

This study presents 1,665 newly sequenced *K. pneumoniae* isolates from a 15-year longitudinal clinical surveillance network (HuiNet). When integrated with 1,498 public genomes, these isolates establish a scalable, pan-genome-based framework for linking genotype to phenotype in bacterial pathogens. We demonstrate that integrating large-scale genomic screens with structural prediction and temporal trend analysis provides functional and evolutionary insight beyond statistical prediction alone. The progressive replacement of *bla*_{CTX-M-14}-associated configurations by the *ISEcp1*–*bla*_{CTX-M-15} transposition unit over 15 years, coupled with coding repeat variation in *ompA*, illustrates the diversity of genomic variation that shapes the resistance landscape. These findings highlight that the mobile-element contexts surrounding resistance genes may themselves be targets of selection, a dimension of AMR evolution that extends beyond the acquisition and spread of individual resistance determinants.

Abbreviations

<i>AMR</i>	Antimicrobial resistance
<i>HuiNet</i>	Anhui Antimicrobial Surveillance Network
<i>AMK</i>	Amikacin
<i>GEN</i>	Gentamicin
<i>TOB</i>	Tobramycin
<i>CAZ</i>	Ceftazidime
<i>CXM</i>	Cefuroxime
<i>CZO</i>	Cefazolin
<i>FOX</i>	Cefoxitin
<i>CIP</i>	Ciprofloxacin
<i>LVX</i>	Levofloxacin
<i>ETP</i>	Ertapenem
<i>IPM</i>	Imipenem
<i>MEM</i>	

Meropenem
SAM
Ampicillin-sulbactam
TZP
Piperacillin-tazobactam
SXT
Trimethoprim-sulfamethoxazole
TCY
Tetracycline
NIT
Nitrofurantoin
CLSI
Clinical and Laboratory Standards Institute
MICs
Minimum inhibitory concentrations
S
Susceptible
R
Resistant
I
Intermediate
MDR
Multidrug resistance
XDR
Extensively drug-resistant
STs
Sequence types
ARGs
Antimicrobial resistance genes
cgSNPs
Core-genome SNPs
LR
Logistic regression
RF
Random forest
SVM
Support vector machine
XGB
Extreme Gradient Boosting
SNPs

Single-nucleotide polymorphisms

AUC

Area under the receiver operating characteristic curve

LOR

Log-odds ratio

SNVs

Single-nucleotide variants

pLDDT

Predicted local distance difference test

KL

K locus

Declarations

Ethics approval and consent to participate

The maintenance and manipulation of all the clinical isolates of *K. pneumoniae* were performed in the biosafety level2 laboratories at the member hospitals of the HuiNet. This research was authorized by the Ethics Committee of the first affiliated hospital of Anhui medical university with approval number of PJ2023-13-38.

Consent for publication

Not applicable.

Competing Interests

The findings described in this manuscript are the subject of a Chinese patent application. The authors declare no other competing interests.

Funding

This work is supported by the National Natural Science Foundation of China (grant nos. 32422019, 32430017, 32125009, 32400509, 82400132, U24A20643), the National Key R&D Program of China (grant nos. 2025YFC3410300, 2022YFC3400300), the Natural Science Foundation of Shaanxi Province (grant no. 2024JC-JCQN-28), the Fundamental Research Funds for the Central Universities (xzy012024088), the Scientific Research Program of Shaanxi Provincial Department of Education (23JK0290), Anhui University Natural Science Research Project (2024AH050826, 2023AH010083, 2024AH010114), Anhui Province Postdoctoral Research Funding Project (2023B700), Anhui Province Clinical Medical Research Transformation Special Project (202304295107020032, 202304295107020043), and Esoteric People of Wan Education [2019] 54.

Author Contribution

X.Y., K.Y. and Y. L conceived and supervised the study. X.Y., H. L. and S. W. designed the model, performed data analyses, and wrote the manuscript. S.W. contributed to the evaluation of the ensemble model. H. L., J.L., H.X. and Y.L. collected the data. S.W., H.X., X.Y., K. Y. and J. L. contributed to data analysis and result interpretation. All authors read and approved the final manuscript.

Acknowledgement

We would like to thank the staff members of the HuiNet Research Network as well as all the authors for making their valuable research data public.

Data Availability

The next-generation sequencing data and nanopore long reads data generated in this study have been deposited in the National Genomic Data Center (<https://ngdc.cncb.ac.cn/gsa/>) under the accession number CRA044521, and the genome assemblies have been deposited in Zenodo under DOI: <https://doi.org/10.5281/zenodo.20713124>. The public obtained data used in this study has been summarized in supplementary materials. The Prodigal annotation results for all 3,163 genomes have been deposited in a separate Zenodo repository under DOI: <https://doi.org/10.5281/zenodo.20713179>. The antimicrobial susceptibility phenotypes of the isolates sequenced in this study are provided in Supplementary Table 2. The source code used in this study is available for download at https://github.com/xjtu-omics/Kpn_pan-genome_of_3163_isolates. The repository is free for noncommercial use by academic, government and nonprofit/not-for-profit institutions.

References

1. Mukherjee S, Mitra S, Dutta S, Basu S. Neonatal Sepsis: The Impact of Carbapenem-Resistant and Hypervirulent *Klebsiella pneumoniae*. *Front Med (Lausanne)*. 2021;8:634349. <https://doi.org/10.3389/fmed.2021.634349>.
2. Jun JB. *Klebsiella pneumoniae* Liver Abscess. *Infect Chemother*. 2018;50:210–8. <https://doi.org/10.3947/ic.2018.50.3.210>.
3. Lam MMC, et al. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat Commun*. 2021;12:4188. <https://doi.org/10.1038/s41467-021-24448-3>.
4. Spadar A, Perdigao J, Campino S, Clark TG. Large-scale genomic analysis of global *Klebsiella pneumoniae* plasmids reveals multiple simultaneous clusters of carbapenem-resistant hypervirulent strains. *Genome Med*. 2023;15:3. <https://doi.org/10.1186/s13073-023-01153-y>.

5. Su M, Satola SW, Read TD. Genome-Based Prediction of Bacterial Antibiotic Resistance. *J Clin Microbiol*. 2019;57. <https://doi.org/10.1128/JCM.01405-18>.
6. Kordova K, Collins C, Parkhill J. Comparative assessment of annotation tools reveals critical antimicrobial resistance knowledge gaps in *Klebsiella pneumoniae*. *Sci Rep*. 2025;15:40495. <https://doi.org/10.1038/s41598-025-24333-9>.
7. Li Y, Kumar S, Zhang L. Mechanisms of Antibiotic Resistance and Developments in Therapeutic Strategies to Combat *Klebsiella pneumoniae* Infection. *Infect Drug Resist*. 2024;17:1107–19. <https://doi.org/10.2147/IDR.S453025>.
8. Agyekum A, et al. Predictability of Phenotype in Relation to Common beta-Lactam Resistance Mechanisms in *Escherichia coli* and *Klebsiella pneumoniae*. *J Clin Microbiol*. 2016;54:1243–50. <https://doi.org/10.1128/JCM.02153-15>.
9. Tsai YK, et al. *Klebsiella pneumoniae* outer membrane porins OmpK35 and OmpK36 play roles in both antimicrobial resistance and virulence. *Antimicrob Agents Chemother*. 2011;55:1485–93. <https://doi.org/10.1128/AAC.01275-10>.
10. Nicolas-Chanoine MH, Mayer N, Guyot K, Dumont E, Pages JM. Interplay Between Membrane Permeability and Enzymatic Barrier Leads to Antibiotic-Dependent Resistance in *Klebsiella Pneumoniae*. *Front Microbiol*. 2018;9:1422. <https://doi.org/10.3389/fmicb.2018.01422>.
11. Jimenez-Castellanos JC, et al. Envelope proteome changes driven by RamA overproduction in *Klebsiella pneumoniae* that enhance acquired beta-lactam resistance. *J Antimicrob Chemother*. 2018;73:88–94. <https://doi.org/10.1093/jac/dkx345>.
12. Pei N, et al. Genome-wide association study of *Klebsiella pneumoniae* identifies variations linked to carbapenems resistance. *Front Microbiol*. 2022;13:997769. <https://doi.org/10.3389/fmicb.2022.997769>.
13. Yang MR, Wu YW. Enhancing predictions of antimicrobial resistance of pathogens by expanding the potential resistance gene repertoire using a pan-genome-based feature selection approach. *BMC Bioinformatics*. 2022;23:131. <https://doi.org/10.1186/s12859-022-04666-2>.
14. Rodrigues C, Lanza VF, Peixe L, Coque TM, Novais A. Phylogenomics of Globally Spread Clonal Groups 14 and 15 of *Klebsiella pneumoniae*. *Microbiol Spectr*. 2023;11:e0339522. <https://doi.org/10.1128/spectrum.03395-22>.
15. Cross BJ, Partridge SR, Sheppard AE. Impacts of mobile genetic elements on antimicrobial resistance genes in gram-negative pathogens: Current insights and genomic approaches. *Microbiol Res*. 2026;302:128340. <https://doi.org/10.1016/j.micres.2025.128340>.
16. Wong A. Epistasis and the Evolution of Antimicrobial Resistance. *Front Microbiol*. 2017;8:246. <https://doi.org/10.3389/fmicb.2017.00246>.
17. Villa L, et al. Diversity, virulence, and antimicrobial resistance of the KPC-producing *Klebsiella pneumoniae* ST307 clone. *Microb Genom*. 2017;3:e000110. <https://doi.org/10.1099/mgen.0.000110>.

18. Shi Q, et al. Epidemiology of carbapenem-resistant *Klebsiella pneumoniae* in China and the evolving trends of predominant clone ST11: a multicentre, genome-based study. *J Antimicrob Chemother.* 2024;79:2292–7. <https://doi.org/10.1093/jac/dkae227>.
19. Poirel L, Decousser JW, Nordmann P. Insertion sequence ISEcp1B is involved in expression and mobilization of a bla(CTX-M) beta-lactamase gene. *Antimicrob Agents Chemother.* 2003;47:2938–45. <https://doi.org/10.1128/AAC.47.9.2938-2945.2003>.
20. Ma L, Siu LK, Lu PL. Effect of spacer sequences between bla(CTX-M) and ISEcp1 on bla(CTX-M) expression. *J Med Microbiol.* 2011;60:1787–92. <https://doi.org/10.1099/jmm.0.033910-0>.
21. Tunstall T, et al. Combining structure and genomics to understand antimicrobial resistance. *Comput Struct Biotechnol J.* 2020;18:3377–94. <https://doi.org/10.1016/j.csbj.2020.10.017>.
22. Magiorakos AP, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect.* 2012;18:268–81. <https://doi.org/10.1111/j.1469-0691.2011.03570.x>.
23. Manni M, Berkeley MR, Seppely M, Simao FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol.* 2021;38:4647–54. <https://doi.org/10.1093/molbev/msab199>.
24. Wick RR, Judd LM, Gorrie CL, Holt KE, Unicycler. Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* 2017;13:e1005595. <https://doi.org/10.1371/journal.pcbi.1005595>.
25. Koren S, et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 2017;27:722–36. <https://doi.org/10.1101/gr.215087.116>.
26. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019;37:540–6. <https://doi.org/10.1038/s41587-019-0072-8>.
27. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17:155–8. <https://doi.org/10.1038/s41592-019-0669-3>.
28. Vaser R, Sovic I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 2017;27:737–46. <https://doi.org/10.1101/gr.214270.116>.
29. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE.* 2014;9:e112963. <https://doi.org/10.1371/journal.pone.0112963>.
30. Hunt M, et al. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* 2015;16:294. <https://doi.org/10.1186/s13059-015-0849-0>.
31. Diancourt L, Passet V, Verhoef J, Grimont PA, Brisse S. Multilocus sequence typing of *Klebsiella pneumoniae* nosocomial isolates. *J Clin Microbiol.* 2005;43:4178–82. <https://doi.org/10.1128/JCM.43.8.4178-4182.2005>.
32. Zankari E, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* 2012;67:2640–4. <https://doi.org/10.1093/jac/dks261>.

33. Carattoli A, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother*. 2014;58:3895–903. <https://doi.org/10.1128/AAC.02412-14>.
34. Ondov BD, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17:132. <https://doi.org/10.1186/s13059-016-0997-x>.
35. Page AJ, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*. 2015;31:3691–3. <https://doi.org/10.1093/bioinformatics/btv421>.
36. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*. 2015;31:3350–2. <https://doi.org/10.1093/bioinformatics/btv383>.
37. Hyatt D, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119. <https://doi.org/10.1186/1471-2105-11-119>.
38. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068–9. <https://doi.org/10.1093/bioinformatics/btu153>.
39. Page AJ et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2, e000056 (2016). <https://doi.org/10.1099/mgen.0.000056>
40. Croucher NJ, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res*. 2015;43:e15. <https://doi.org/10.1093/nar/gku1196>.
41. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312–3. <https://doi.org/10.1093/bioinformatics/btu033>.
42. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res*. 2024;52:W78–82. <https://doi.org/10.1093/nar/gkae268>.
43. Tonkin-Hill G, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*. 2020;21. <https://doi.org/10.1186/s13059-020-02090-4>.
44. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013). <https://doi.org/10.48550/arXiv.1303.3997>
45. Danecek P et al. Twelve years of SAMtools and BCFtools. *Gigascience* 10 (2021). <https://doi.org/10.1093/gigascience/giab008>
46. Garrison E, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*. 2018;36:875–9. <https://doi.org/10.1038/nbt.4227>.
47. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6:80–92. <https://doi.org/10.4161/fly.19695>.
48. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res*. 2025;53:D672–7. <https://doi.org/10.1093/nar/gkae909>.
49. Pedregosa F, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–30.

50. Chen TQ, Guestrin C, XGBoost: A Scalable Tree Boosting System. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
51. Alcock BP, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48:D517–25. <https://doi.org/10.1093/nar/gkz935>.
52. Feldgarden M, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep.* 2021;11:12728. <https://doi.org/10.1038/s41598-021-91456-0>.
53. Johnson M, et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* 2008;36:W5–9. <https://doi.org/10.1093/nar/gkn201>.
54. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>.
55. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–9. <https://doi.org/10.1038/s41586-021-03819-2>.
56. Ittisoponpisan S, et al. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol.* 2019;431:2197–212. <https://doi.org/10.1016/j.jmb.2019.04.009>.
57. gggenes. Draw Gene Arrow Maps in 'ggplot2' v. 0.6.0. CRAN; 2025.
58. Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2009.
59. Armitage P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics.* 1955;11. <https://doi.org/10.2307/3001775>.
60. Liang W, et al. Anti-Restriction Protein, KlcA(HS), Promotes Dissemination of Carbapenem Resistance. *Front Cell Infect Microbiol.* 2017;7:150. <https://doi.org/10.3389/fcimb.2017.00150>.
61. Liang W, et al. KlcAHS genes are ubiquitous in clinical, blaKPC-2-positive, *Klebsiella pneumoniae* isolates. *Infect Genet Evol.* 2019;70:84–9. <https://doi.org/10.1016/j.meegid.2019.02.021>.
62. Heng H, et al. Profiling the landscape of carbapenem resistance and hypervirulence in *Klebsiella pneumoniae*: A global epidemiological analysis of the plasmidome. *Drug Resist Updat.* 2025;81:101254. <https://doi.org/10.1016/j.drug.2025.101254>.
63. Llobet E, March C, Gimenez P, Bengoechea JA. *Klebsiella pneumoniae* OmpA confers resistance to antimicrobial peptides. *Antimicrob Agents Chemother.* 2009;53:298–302. <https://doi.org/10.1128/AAC.00657-08>.
64. Ortiz-Suarez ML, Samsudin F, Piggot TJ, Bond PJ, Khalid S. Full-Length OmpA: Structure, Function, and Membrane Interactions Predicted by Molecular Dynamics Simulations. *Biophys J.* 2016;111:1692–702. <https://doi.org/10.1016/j.bpj.2016.09.009>.
65. Confer AW, Ayalew S. The OmpA family of proteins: roles in bacterial pathogenesis and immunity. *Vet Microbiol.* 2013;163:207–22. <https://doi.org/10.1016/j.vetmic.2012.08.019>.

66. Poirel L, Lartigue MF, Decousser JW, Nordmann P. ISEcp1B-mediated transposition of blaCTX-M in *Escherichia coli*. *Antimicrob Agents Chemother*. 2005;49:447–50.
<https://doi.org/10.1128/AAC.49.1.447-450.2005>.
67. Abbassi MS, et al. Genetic characterisation of CTX-M-15-producing *Klebsiella pneumoniae* and *Escherichia coli* strains isolated from stem cell transplant patients in Tunisia. *Int J Antimicrob Agents*. 2008;32:308–14. <https://doi.org/10.1016/j.ijantimicag.2008.04.009>.
68. Mallawaarachchi S, et al. Genome-wide association, prediction and heritability in bacteria with application to *Streptococcus pneumoniae*. *NAR Genom Bioinform*. 2022;4:lqac011.
<https://doi.org/10.1093/nargab/lqac011>.
69. James T, Williamson B, Tino P, Wheeler N. Whole-genome phenotype prediction with machine learning: open problems in bacterial genomics. *Bioinformatics*. 2025;41.
<https://doi.org/10.1093/bioinformatics/btaf206>.
70. Saber MM, Shapiro BJ. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom*. 2020;6.
<https://doi.org/10.1099/mgen.0.000337>.
71. Wyres KL, et al. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of *Klebsiella pneumoniae*. *PLoS Genet*. 2019;15:e1008114.
<https://doi.org/10.1371/journal.pgen.1008114>.
72. Wyres KL, Holt KE. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Curr Opin Microbiol*. 2018;45:131–9.
<https://doi.org/10.1016/j.mib.2018.04.004>.
73. Pedersen T et al. Horizontal Plasmid Transfer among *Klebsiella pneumoniae* Isolates Is the Key Factor for Dissemination of Extended-Spectrum beta-Lactamases among Children in Tanzania. *mSphere* 5 (2020). <https://doi.org/10.1128/mSphere.00428-20>
74. Hyun JC, Monk JM, Szubin R, Hefner Y, Palsson BO. Global pathogenomic analysis identifies known and candidate genetic antimicrobial resistance determinants in twelve species. *Nat Commun*. 2023;14:7690. <https://doi.org/10.1038/s41467-023-43549-9>.
75. Zhou K, Aertsen A, Michiels CW. The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol Rev*. 2014;38:119–41. <https://doi.org/10.1111/1574-6976.12036>.
76. Mathers AJ, Peirano G, Pitout JD. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clin Microbiol Rev*. 2015;28:565–91. <https://doi.org/10.1128/CMR.00116-14>.
77. Ikhimiukor OO, et al. Clonal background and routes of plasmid transmission underlie antimicrobial resistance features of bloodstream *Klebsiella pneumoniae*. *Nat Commun*. 2024;15:6969.
<https://doi.org/10.1038/s41467-024-51374-x>.
78. Additional. file information.
79. Additional file 1. Supplementary figures S1-S15.
80. Additional file 2. Supplementary Tables S1-S32.

Figures

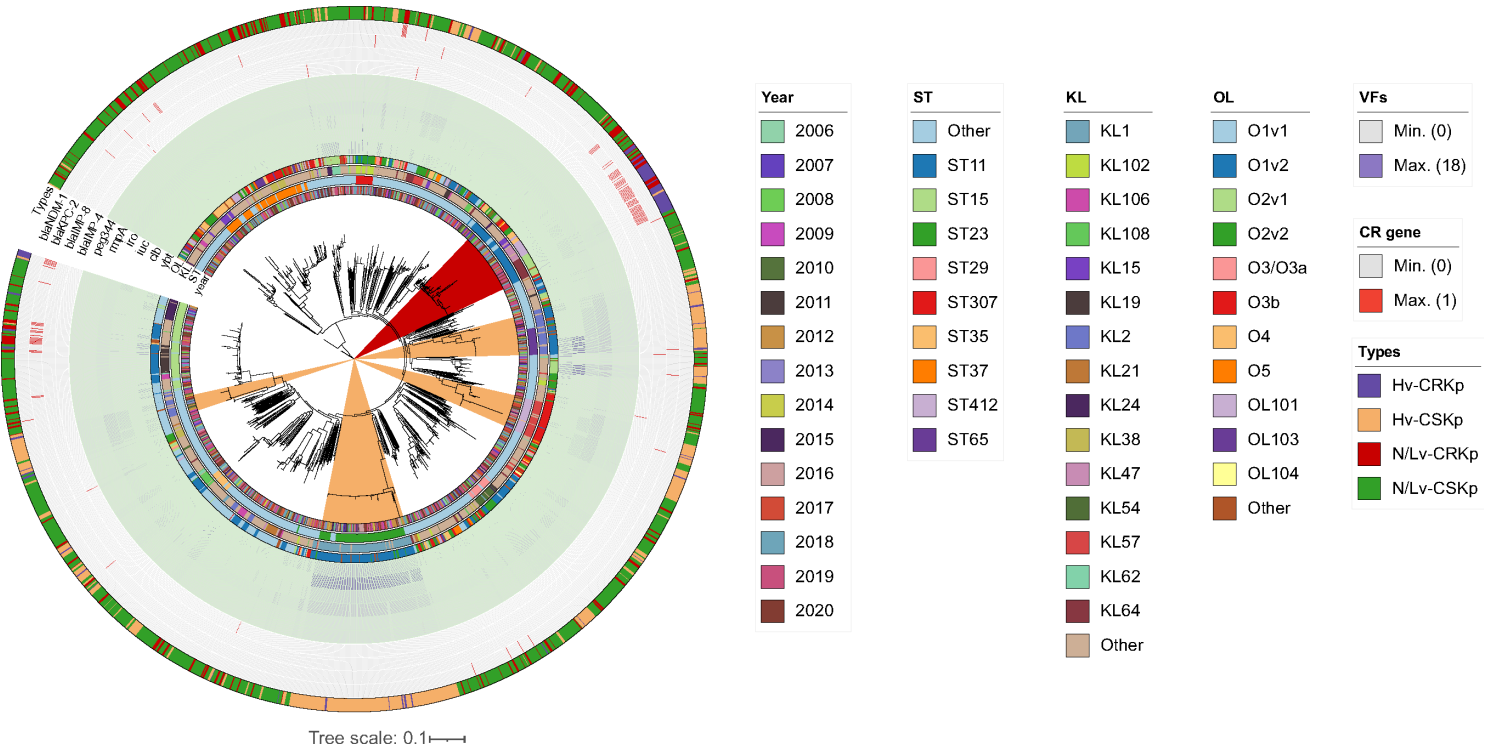


Figure 1

Characterization of 1,665 *K. pneumoniae* isolates from HuiNet. Phylogenetic tree of 1665 *K. pneumoniae* isolates. The color of the innermost circle represented the isolation year ranging from 2006 to 2020. Nine major ST types were displayed and remaining ST types were belonged to other. The heatmap of carbapenem resistance genes indicated blaDNM-1, blaKPC-2, blaIMP-4 and blaIMP-8 to be given beside the relevant phylogeny. The number of core and plasmid-mediated virulence factors in relevant phylogeny was represented by the outermost heatmap. Clusters of high-virulence-only and carbapenem-resistant isolates were highlighted on the inner tree in orange-yellow and red, respectively.

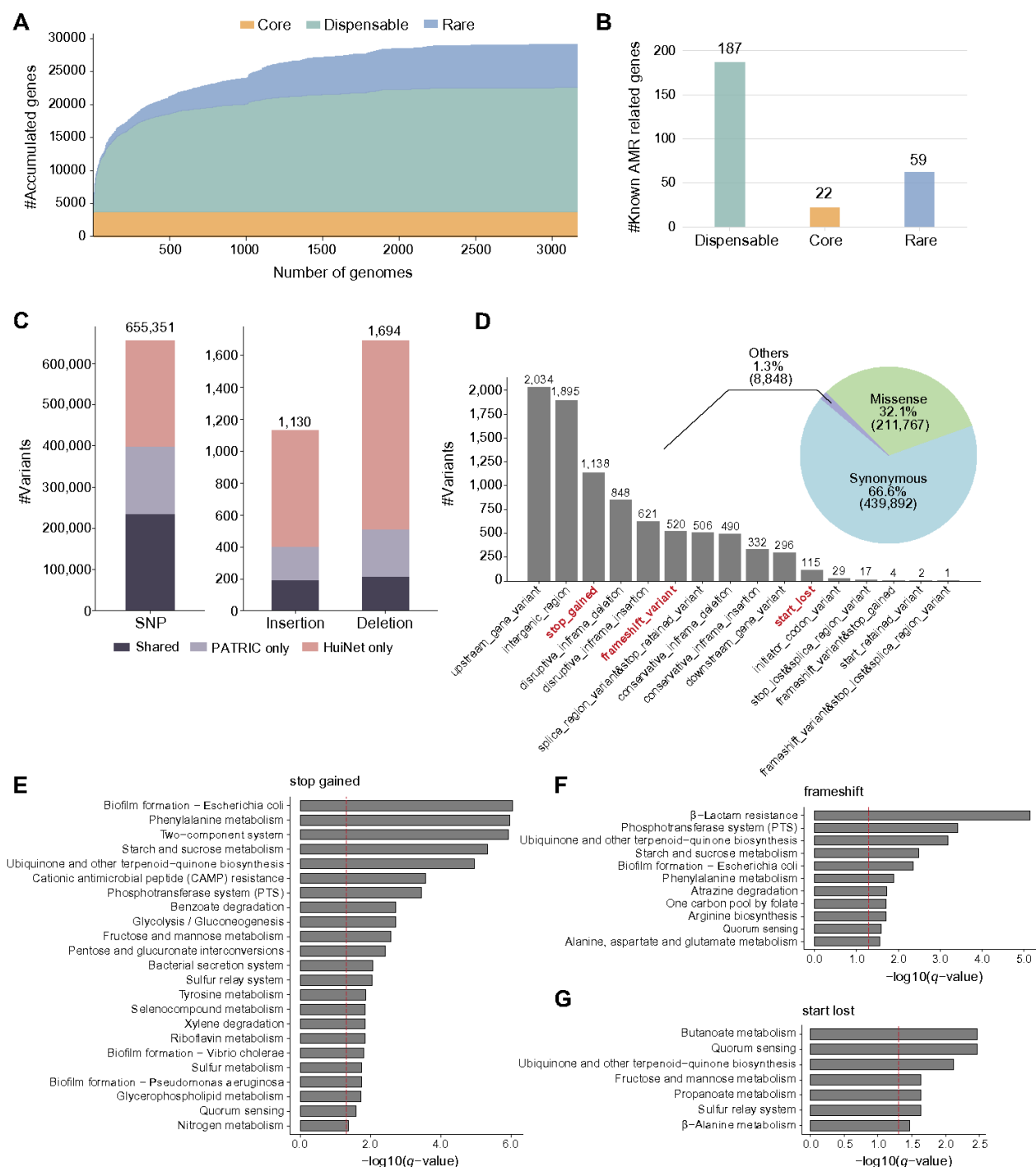


Figure 2

Core-gene variants analysis of 3,163 *K. pneumoniae* genomes. **A**, Gene accumulation curves for core genes (present in >90% of genomes), dispensable genes (present in >3 and <90% of genomes), and rare genes (present in ≤3 genomes). **B**, Distribution of known antimicrobial resistance (AMR) genes across pan-genome frequency classes. **C**, Composition of core gene variants identified from the integrated HuiNet and PATRIC dataset. Counts of single-nucleotide polymorphisms (SNPs), insertions, and

deletions, partitioned into variants shared between datasets and those unique to PATRIC or HuiNet. **D**, Functional annotation of core-gene variants into synonymous, missense, and other mutation types. **E-G**, KEGG pathway enrichment of genes harbouring high-impact core-gene variants of Frameshift (**E**), Start-loss (**F**), and Stop-gain variants (**G**).

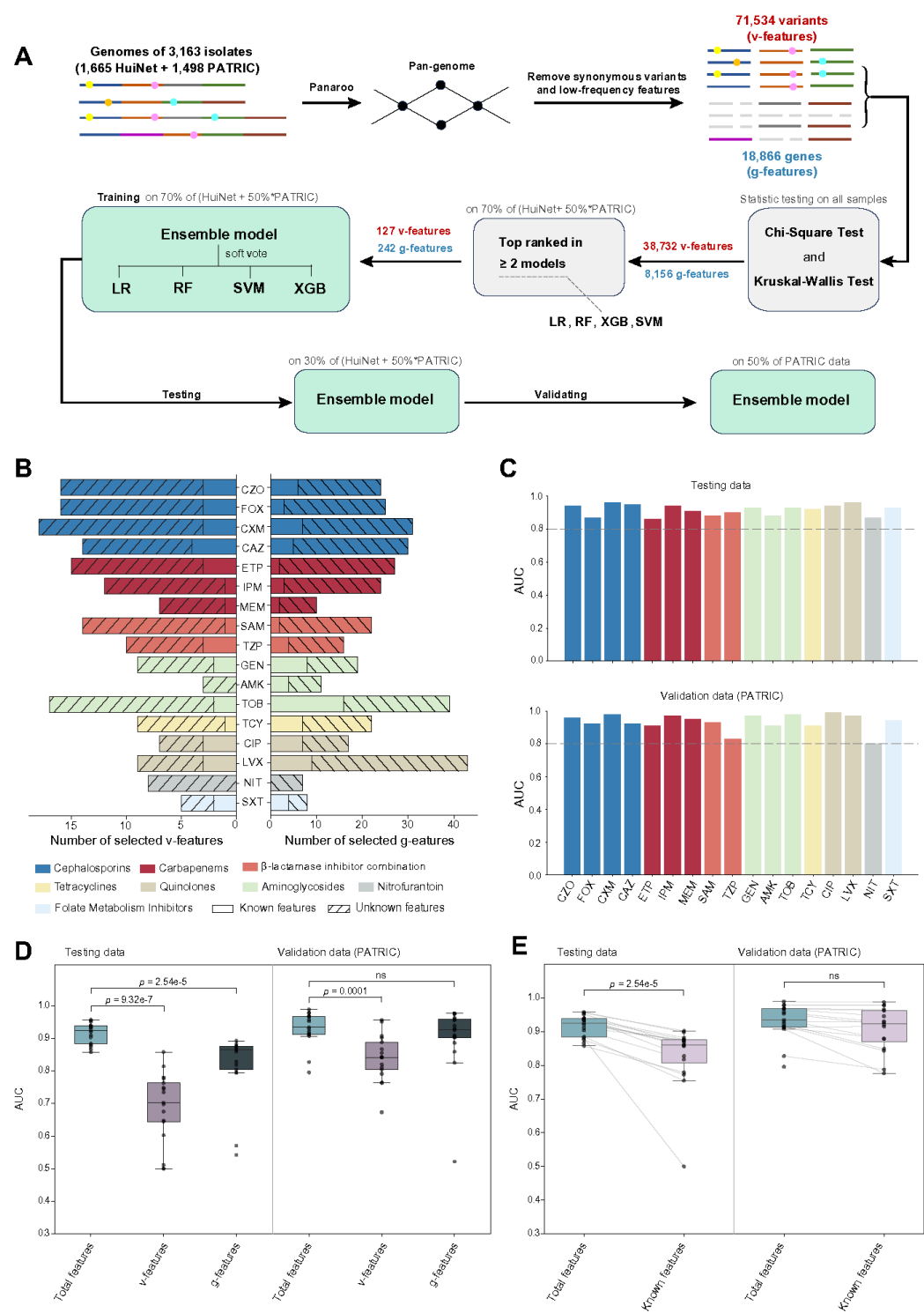


Figure 3

AMR genomic features for 17 antibiotics. A, Schematic of the feature discovery and model construction pipeline. From the pan-genome of 3,163 isolates, dispensable gene presence/absence (g-features) and non-synonymous core gene variants (v-features) were extracted. A two-stage selection combining statistical association testing and multi model machine learning ranking yielded a final set of 127 v-features and 242 g-features. Ensemble models were trained, tested, and externally validated on independent datasets. **B,** Distribution of selected v-features and g-features across antibiotic classes. **C,** Area under the receiver operating characteristic curve (AUC) for each of the 17 antibiotic-specific models on the internal test set and the external PATRIC validation set. **D,** Performance comparison among models built with the full feature set, v-features alone, and g-features alone. p-values were calculated by one-sided Mann-Whitney U test; ns, not significant. **E,** Performance comparison between models using all 369 selected features and models restricted to the 46 previously known ARGs features.

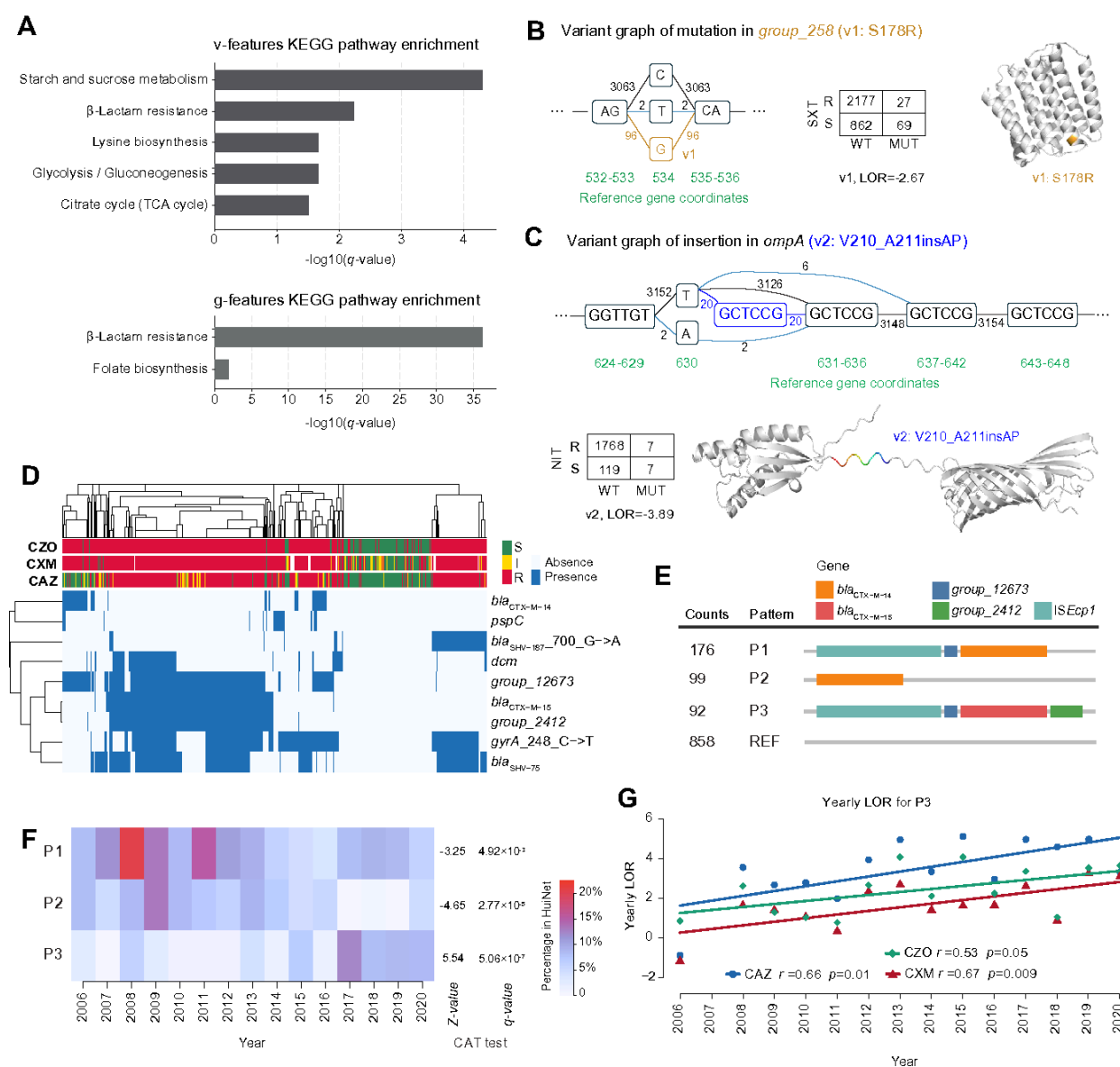


Figure 4

Mechanisms of AMR genomic features. **A**, KEGG pathway enrichment analysis for genes harbouring v-features and genes represented by g-features. Significantly enriched (q -value < 0.05) pathways are shown, with β -lactam resistance as a shared significantly enriched signal. **B-C**, Structural and population-level characterization of two prioritized v-features: (**B**) *group_258* S178R missense variant associated with sulfamethoxazole-trimethoprim (SXT), and (**C**) hexanucleotide repeat expansion (GCTCCG) in *ompA* associated with nitrofurantoin (NIT). For each variant, the predicted protein structure with the location of the variation highlighted, the pan-genome graph representation, and a 2×2 contingency table of resistant (R) versus susceptible (S) phenotypes against variant presence (MUT) and absence (WT) are displayed, together with the corresponding log-odds ratio (LOR). **D**, Clustered heatmap based on the presence/absence matrix of the union of the top five features for cefuroxime (CXM), cefazolin (CZO), and ceftazidime (CAZ). Isolates were clustered using the Ward.D2 method, showing clear separation of R and S phenotypes for each cephalosporin. **E**, Schematic of the three most frequent gene-order patterns (P1–P3) observed in HuiNet isolates, defined by the relative arrangement of *ISEcp1*, *bla*_{CTX-M-15}, *bla*_{CTX-M-14}, *group_2412*, and *group_12673*. **F**, Yearly prevalence of patterns P1–P3 among HuiNet isolates (2006–2020). P3 showed a significant temporal increase (Cochran-Armitage trend test, Z -value = 5.54, $q = 1.74 \times 10^{-7}$), whereas P1 and P2 declined significantly (Z -values = -3.25 and -4.65, $q = 4.92 \times 10^{-3}$ and 2.77×10^{-5} , respectively). **G**, Correlation between collection year and the yearly LOR of pattern P3 for CAZ, CXM, and CZO in the HuiNet cohort. Each point represents the LOR for a year in which P3 was detected; solid lines show linear regressions with Pearson correlation coefficients ($r = 0.66$, 0.67 , and 0.53 ; all $p \leq 0.05$). The positive correlations indicate that P3 became increasingly associated with cephalosporin resistance over time.

Supplementary Files

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

- [Additionalfile1.docx](#)
- [Additionalfile2.xlsx](#)