Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genomic and plasmid-level identity between an environmental Klebsiella pneumoniae isolate from India and the high-risk clinical clone HS11286: a comparative genomic analysis of the K. pneumoniae species complex

Noora Ayesha
Yenepoya University
Sree Sankar Darveekaran Nair
sreesankar162@gmail.com

Yenepoya University

Research Article

Keywords: Klebsiella pneumoniae, environmental reservoir, comparative genomics, average nucleotide identity, plasmid, antimicrobial resistance

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11057485/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 is a leading cause of multidrug-resistant infection, and its persistence in environmental reservoirs is increasingly recognized as a One Health concern.

Objectives

To characterize the genomic relatedness of clinical and environmental K. pneumoniae species complex isolates from India to the internationally significant high-risk clinical strain HS11286.

Methods

Eleven publicly available *K. pneumoniae* species complex genomes from clinical, river, lake, raw milk, and soil sources were typed using Kleborate and compared by pan-genome (Panaroo) and core-genome phylogenetic analysis (IQ-TREE), average nucleotide identity (fastANI), plasmid replicon typing (PlasmidFinder), and accessory-genome Jaccard similarity.

Results

One river water isolate, DCP2, shared sequence type ST11, capsule locus KL103 and O-locus OL2α.1 with HS11286, with a whole-genome average nucleotide identity of 99.99%, the highest pairwise value in the dataset. The two isolates formed a fully supported sister pair (bootstrap = 100) with near-zero branch length, shared an identical complement of five plasmid replicon types at matching sequence identity, and carried the same nine resistance gene classes, including KPC-2 and rmtB. Accessory-genome Jaccard similarity between HS11286 and DCP2 was 0.9956, with 436 accessory genes shared exclusively by this pair.

Conclusions

DCP2 and HS11286 represent the same clonal lineage, sharing near-identical core, accessory and plasmid genomic content despite different countries and years of isolation, providing genomic evidence that environmental water can harbour, and potentially disseminate, high-risk clinical K. pneumoniae clones.

1. Background

Klebsiella pneumoniae is a Gram-negative, encapsulated member of the Enterobacteriaceae and a leading cause of healthcare-associated pneumonia, bloodstream infection, urinary tract infection, and liver abscess, with particular concern attached to carbapenem-resistant and hypervirulent lineages [1–3]. A small number of globally disseminated sequence types, including ST11, ST15, ST147, ST258, and ST307, account for a disproportionate share of high-risk, multidrug-resistant clinical isolates worldwide [1, 3]. HS11286, a carbapenemase-producing ST11 strain isolated from human sputum in Shanghai, China in 2011, is among the most extensively characterized of these reference genomes and carries a well-defined complement of six plasmids alongside its chromosome [4].

Alongside its clinical burden, K. pneumoniae is increasingly recognized as an environmentally competent organism, with reports describing high-risk clinical lineages and the mobile resistance elements they carry, recovered from river water, wastewater, and other non-clinical sources [2, 5]. This has been framed within a One Health perspective, in which environmental compartments are treated as potential reservoirs and conduits for the persistence and spread of clinically significant antimicrobial resistance, rather than as epidemiologically separate from the clinic [5].

Large-scale comparative genomic surveys have characterized global pan-genome structure and phylogenetic diversity in K. pneumoniae using broad, geographically representative genome sets [3]. Such country-representative approaches necessarily sample sparsely within any one country and do not resolve fine-grained, sequence-type-level relationships between individual clinical and environmental isolates; the authors of one recent global pan-genome and phylogenetic survey explicitly noted this as a limitation of their approach and recommended that future work integrate multilocus sequence typing and clonal lineage analysis at finer resolution [3]. Individual reports of specific environmental K. pneumoniae genomes from India exist, including single-isolate characterizations of an ST147 strain from a river confluence and of an isolate deposited from an unspecified water source [2, 6], but a systematic, multi-source comparison of Indian environmental and clinical K. pneumoniae species complex genomes against a well-characterized international clinical reference has not, to our knowledge, been reported.

The present study addresses this gap. We compiled a set of publicly available K. pneumoniae species complex genomes recovered from clinical, river, lake, raw milk and soil sources, alongside the clinical reference strain HS11286, and asked: (i) which sequence types and capsule/O-antigen types are represented across these sources; (ii) how closely related, at the whole-genome level, are the environmental isolates to a well-characterised high-risk clinical strain; and (iii) do the isolates most closely related to the clinical reference also carry a comparable resistance and virulence gene complement, including on shared mobile genetic elements.

2. Materials and Methods

2.1 Genome dataset

Eleven publicly available K. pneumoniae species complex genome assemblies were retrieved from the National Center for Biotechnology Information (NCBI) RefSeq database, spanning clinical, river water, lake water, raw milk and soil sources (**Table 1**). The clinical reference strain HS11286 (GCF_000240185.1), a complete, closed genome, was included as a well-characterized international comparator [4]. Isolation source and geographic provenance for each genome were confirmed against the corresponding NCBI BioSample record prior to analysis.

2.2 Species identification, sequence typing and capsule/O typing

Species assignment, multilocus sequence type (ST), capsule (K) locus, O-antigen locus, and composite virulence and resistance scores were determined for all genomes using Kleborate (K. pneumoniae species complex module) [7]. Kleborate does not type organisms outside the K. pneumoniae species complex; the two most distantly related genomes in the dataset (*K. ornithinolytica* and *K. aerogenes*) were therefore included in phylogenetic and average nucleotide identity analyses as outgroups but were not assigned ST/K/O designations.

2.3 Pan-genome and phylogenetic analysis

Orthologous gene clustering and pan-genome composition (core, soft-core, shell, and cloud gene categories) were determined using Panaroo [8]. The resulting core-genome alignment was used to infer a maximum-likelihood phylogeny in IQ-TREE [9], with branch support assessed by ultrafast bootstrap approximation.

2.4 Average nucleotide identity

Whole-genome average nucleotide identity (ANI) was calculated for all pairwise genome combinations using fastANI [10]. Values ≥ 95% were interpreted as consistent with membership of the same species, and values ≥ 99.5% as consistent with membership of the same clonal lineage, following conventional thresholds used in bacterial genomic epidemiology [1, 10].

2.5 Antimicrobial resistance and virulence gene content

Antimicrobial resistance determinants were identified against the Comprehensive Antibiotic Resistance Database (CARD) [11]and virulence-associated genes were identified using the Virulence Factor Database (VFDB) [12]. Gene presence–absence across all eleven genomes was visualized as heatmaps using pheatmap (R). Genes were not hierarchically clustered; rows were ordered first by conservation category core (present in all 11 genomes), accessory (present in 2–10 genomes), or unique

(present in exactly one genome), and second by functional category, assigned from gene nomenclature (e.g. fim* genes as type I fimbrial, mrk* genes as type III fimbrial, ybt/ent/fep/iro genes as iron acquisition). Genomes (columns) were displayed in a fixed order and were not clustered.

2.6 Plasmid replicon typing

Plasmid replicon content was determined for HS11286 and for the most closely related environmental isolate (DCP2, identified by ANI analysis) using PlasmidFinder (*Enterobacteriales* database) [13], which identifies replicon marker sequences by alignment against a curated reference database and reports percent nucleotide identity and alignment coverage for each hit.

2.7 Accessory-genome Jaccard similarity

As an additional independent measure of pairwise relatedness, Jaccard similarity indices were calculated for each genome pair directly from the Panaroo gene presence–absence matrix using a custom R script. For each pair, a whole-pan-genome Jaccard index (genes shared divided by genes present in either genome) and an accessory-genome-only Jaccard index (restricted to genes not present in all eleven genomes, i.e., excluding the trivially shared core genome) were calculated, along with the number of accessory genes present in that pair and no other genome in the dataset.

3. Results

3.1 Genome dataset and assembly quality

The eleven genomes comprised one clinical reference (HS11286), five water isolates (three river, one lake, one further river isolate), one raw milk isolate, and four soil isolates, the latter including two genomes falling outside the *K. pneumoniae* species complex (*K. ornithinolytica*, *K. aerogenes*) (**Table 1**). Assembly contiguity varied considerably across the dataset, from single-contig, chromosome-level assemblies (N50 > 5.1 Mb) to highly fragmented drafts (N50 as low as 82 kb); BUSCO completeness nonetheless remained ≥ 96% for all genomes, indicating that fragmentation did not compromise gene-content completeness assessment.

Table 1. Genome dataset, isolation source, and assembly quality metrics for the eleven genomes analysed. N50 is the contig length above which 50% of the total assembly length is contained, a standard measure of assembly contiguity (higher = less fragmented). BUSCO completeness (%) indicates the proportion of expected single-copy orthologous genes recovered intact (bacteria_odb10 lineage dataset).

Sample type	Strain / label	Country of origin	Accession (GCF)	Species	Size (Mb)	Contigs	N50 (bp)	BUSCO complete (%)
Clinical (reference)	HS11286	China	000240185.1	*K. pneumoniae*	5.68	7	5,333,942	98.4
Water (river)	DCP2	India	039414255.1	*K. pneumoniae*	5.68	4	5,333,348	98.4
Water (river)	IITR008	India	051013965.1	*K. pneumoniae*	5.74	11	5,457,951	98.4
Water (lake, fish-associated)	257	India	019336355.1	*K. pneumoniae*	5.41	48	462,118	98.4
Water (river)	Kp-102	India	035780075.1	*K. pneumoniae*	5.78	54	410,874	98.4
Milk (raw)	KPNCDS21	India	041679105.1	*K. pneumoniae*	5.53	85	227,165	98.4
Soil (rhizosphere)	AWD5	India	001876675.1	*K. pneumoniae*	4.81	132	82,163	96.0
Soil	PSRZ2	India	056171875.1	*K. variicola*	5.69	32	475,786	98.4
Soil	SGM81	India	900187335.1	*K. quasipneumoniae*	5.19	1	5,186,362	96.8
Soil	MGG-12mu	India	035787015.1	*K. ornithinolytica*	5.44	18	807,871	98.4
Soil (agricultural)	G3_AM	India	017742775.1	*K. aerogenes*	5.28	1	5,275,806	98.4

**BUSCO missing (%) is the complement of BUSCO complete and is omitted for brevity; all genomes had < 4.1% missing BUSCOs.*

3.2 Species, sequence type and capsule/O-locus distribution

Kleborate assigned nine of the eleven genomes to the K. pneumoniae species complex; two genomes (PSRZ2, SGM81) were assigned to *K. variicola* and *K. quasipneumoniae*, respectively, and two genomes fell outside the complex entirely (*K. ornithinolytica*, *K. aerogenes*) and were not typed. Among the nine typed *K. pneumoniae genomes*, six distinct sequence types were identified. Two genomes, HS11286 and DCP2 (a river water isolate), shared an identical sequence type (ST11), capsule locus (KL103), and O-locus (OL2α.1), and both returned the highest composite scores in the dataset (virulence score 1, resistance score 3, nine resistance classes) (**Table 2**). A second river isolate, IITR008, carried the internationally recognized high-risk lineage ST147 (capsule locus KL64), consistent with its prior independent characterization as an ST147 environmental isolate [11].

Table 2. Kleborate sequence type, capsule/O-locus and resistance/virulence score summary. Kleborate multilocus sequence type (ST), capsule (K) and O-antigen locus assignments, and composite resistance/virulence scores for the nine genomes within the K. pneumoniae species complex (the two genomes outside the complex, K. ornithinolytica and K. aerogenes, are not typed by Kleborate and are shown as n/a). Virulence score (0–5) reflects the presence of key acquired virulence loci (yersiniabactin, colibactin, aerobactin); resistance score (0–3) reflects acquired resistance tier, with 3 indicating co-occurrence of a carbapenemase and a colistin-resistance determinant, the most clinically severe tier Kleborate assigns.

Strain / label	ST	K locus	O locus	Virulence score	Resistance score	No. resistance genes
HS11286	ST11	KL103	OL2α.1	1	3	13
DCP2	ST11	KL103	OL2α.1	1	3	12
IITR008	ST147	KL64	OL2α.1	1	1	16
Kp-102	ST4467	KL148	OL2α.2	0	1	8
257 (KpS2A9)	ST307	KL102	OL2α.2	0	1	9
KPNCDS21	ST2791-1LV	KL49	OL2α.1	0	0	0
Rhizosphere soil isolate	ST200	KL50	OL3γ	0	0	0
PSRZ2 (*K. variicola*)	ST3936	KL60	OL3α/β	0	0	0
SGM81 (*K. quasipneumoniae*)	ST4671	KL11	OL3α/β	0	0	0
MGG-12mu (*K. ornithinolytica*)	n/a¹	n/a	n/a	n/a	n/a	n/a
G3_AM (*K. aerogenes*)	n/a¹	n/a	n/a	n/a	n/a	n/a

**Not assignable: genome falls outside the K. pneumoniae species complex and is not typed by Kleborate.*

3.3 Whole-genome average nucleotide identity

Pairwise fastANI comparison against the HS11286 reference showed a clear separation between the *K. pneumoniae* species complex genomes (ANI 98.8–99.99%) and the more distantly related complex members and outgroups (ANI 83.7–94.7%), consistent with the species-level boundaries inferred by Kleborate (**Table 3**). Within the *K. pneumoniae* genomes, one isolate stood out clearly from the rest of the dataset: DCP2, a river water isolate, returned an ANI of 99.9932% to HS11286, more than 0.7 percentage points higher than any other genome in the dataset, and comfortably above the 99.5% threshold conventionally used to indicate membership of the same clonal lineage. All remaining *K. pneumoniae* genomes, including the second river isolate IITR008 (ST147), clustered in a narrower 98.8–99.3% band, indicating that they are congeneric but represent genomically distinct strains rather than the same clone.

Table 3. Whole-genome average nucleotide identity (fastANI) of each genome relative to HS11286. Whole-genome average nucleotide identity (ANI), calculated using fastANI, of each genome relative to the clinical reference strain HS11286. Values ≥95% are conventionally interpreted as indicating membership of the same species; values ≥99.5% are conventionally interpreted as indicating membership of the same clonal lineage.

Strain / label	Sample type	Species	ANI to HS11286 (%)
DCP2	Water (river)	*K. pneumoniae*	99.9932
Rhizosphere soil isolate	Soil	*K. pneumoniae*	99.2542
257 (KpS2A9)	Water (lake)	*K. pneumoniae*	99.0792
KPNCDS21	Milk	*K. pneumoniae*	98.9560
Kp-102	Water (river)	*K. pneumoniae*	98.8598
IITR008	Water (river)	*K. pneumoniae*	98.8486
PSRZ2	Soil	*K. variicola*	94.7296
SGM81	Soil	*K. quasipneumoniae*	93.7386
G3_AM	Soil	*K. aerogenes*	86.2849
MGG-12mu	Soil	*K. ornithinolytica*	83.7462

3.4 Core-genome phylogeny

The core-genome alignment comprised 2,417 core genes (present in ≥ 99% of genomes), out of a total pan-genome of 10,720 gene clusters (3,871 shell genes, present in 15–95% of genomes; 4,432 cloud genes, present in < 15% of genomes). This open pan-genome structure, with a core gene fraction of approximately 23%, is consistent with the extensive plasticity of the accessory genome previously reported for this species complex.

The maximum-likelihood core-genome phylogeny (Fig. 1) placed HS11286 and DCP2 as sister taxa with maximal bootstrap support (100) and a near-zero branch length separating them (0.00003 substitutions/site), making this pair the two most closely related tips in the entire tree by a wide margin. This clade grouped with the rhizosphere soil isolate at full support (100), and this three-taxon group in turn joined a clade comprising the lake isolate and IITR008 at comparatively weak support (bootstrap = 45); this deeper node should therefore be interpreted with caution, whereas the HS11286–DCP2 relationship itself is supported unambiguously. The raw milk isolate (KPNCDS21) and the Ganga river isolate (Kp-102) formed a separate, fully supported (100) sister pair. *K. quasipneumoniae* (SGM81) branched basally to the *K. pneumoniae* sensu stricto clade, with *K. variicola* (PSRZ2) and the *K. ornithinolytica* / *K. aerogenes* pair forming more distant outgroups; the latter pair, while well supported (100), was separated from the rest of the tree by substantially longer branch lengths, consistent with their greater evolutionary distance from the *K. pneumoniae* species complex rather than close phylogenetic affinity to one another.

Antimicrobial resistance and virulence gene content

HS11286 and DCP2 carried a near-identical resistance gene complement across the same nine resistance classes (β-lactam, aminoglycoside, fluoroquinolone, fosfomycin, macrolide, phenicol, rifamycin, sulfonamide/trimethoprim and tetracycline), including the carbapenemase gene KPC-2 and the 16S rRNA methyltransferase rmtB, which together confer resistance to carbapenems and essentially all clinically used aminoglycosides (Fig. 2). HS11286 carried one additional resistance gene relative to DCP2 (13 v/s 12; **Table 2**), consistent with the two genomes representing the same lineage with minor, expected strain-level variation rather than identical isolates. In contrast, all other environmental and food-source isolates carried substantially fewer resistance genes and classes, or none (**Table 2**).

Virulence gene content followed a broadly similar pattern of conservation between HS11286 and DCP2 across the type I and type III fimbrial operons, the yersiniabactin (ybt) and enterobactin/salmochelin iron-acquisition loci, and the type VI secretion system gene clusters characteristic of the K. pneumoniae species complex core virulence repertoire (Fig. 3).

3.5 Plasmid replicon content of HS11286 and DCP2

To assess whether the resistance genotype shared by HS11286 and DCP2 was associated with shared mobile genetic elements, both genomes were screened for plasmid replicon markers using PlasmidFinder. Both genomes returned an identical complement of five replicon types containing IncFIB(pKPHS1), IncC, Col(KPHS6), IncR and IncFII(K), at matching percent identity and full-length coverage in every case (**Table 4**). For the two replicons showing < 100% identity to the reference database (IncFII(K), 98.65%; IncR, 99.2%), the underlying nucleotide mismatches occurred at the same alignment positions in both HS11286 and DCP2, indicating that the two genomes carry sequence-identical replicon variants rather than merely the same Inc-type family. In HS11286, a closed, complete genome, all five replicon hits mapped to previously named plasmids (pKPHS1, pKPHS2, pKPHS3 and associated replicons) [1]; in DCP2, a draft assembly, the corresponding hits mapped to three separate contigs, consistent with these replicons residing on extrachromosomal elements in both genomes.

Table 4. **Plasmid replicon content of HS11286 and DCP2, determined by PlasmidFinder against the Enterobacteriales reference database.** Each row is a distinct replicon type (Inc- or Col-type), the sequence a plasmid uses to initiate its own replication and a standard marker for plasmid classification; matching replicon types at matching percent identity in both genomes indicate the two genomes carry sequence-identical, or near-identical, plasmid backbones.

Plasmid replicon	Reference accession	HS11286 identity (%)	DCP2 identity (%)	Match length (bp)
IncFIB(pKPHS1)	CP003223	100.00	100.00	560 / 560
IncC	JN157804	100.00	100.00	417 / 417
Col(KPHS6)	NC_016841	100.00	100.00	178 / 178
IncR	DQ449578	99.20	99.20	251 / 251
IncFII(K)	CP000648	98.65	98.65	148 / 148

3.6 Accessory-genome Jaccard similarity

As an independent, gene-content-based measure of relatedness alongside ANI, phylogeny, and plasmid typing, pairwise Jaccard similarity was calculated for every genome pair from the Panaroo gene presence–absence matrix. HS11286 and DCP2 shared 5,365 of a possible 5,378 gene clusters present in either genome (whole-pan-genome Jaccard = 0.9976); restricting this comparison to the accessory genome alone, the more informative measure, since core genes are shared genome-wide by definition, gave an accessory Jaccard index of 0.9956 (2,948 of 2,961 accessory gene clusters shared).

This degree of accessory-genome sharing was substantially higher than for any other genome pair in the dataset, including the two other sister-taxon pairs that also received maximal (100) bootstrap support in the core-genome phylogeny (**Table 5**): KPNCDS21 (raw milk) and Kp-102 (Ganga river) shared an accessory Jaccard of 0.6972, and the lake isolate (257/KpS2A9) and IITR008 shared 0.6956, both around 0.70, roughly two-thirds the value observed for HS11286–DCP2. The distinction was more pronounced when considering genes shared exclusively by a given pair and absent from all nine remaining genomes: HS11286 and DCP2 shared 436 such genes, compared with 56 and 21 for the next-closest pairs respectively, and zero for HS11286 paired with either of the other two river isolates in the dataset (IITR008 or Kp-102). This substantial, exclusively shared accessory gene pool provides a fourth, independent line of evidence, alongside whole-genome ANI, core-genome phylogenetic clustering, and plasmid replicon identity, that HS11286 and DCP2 share very recent common ancestry, rather than convergent similarity driven by shared ecological niche or geography.

Table 5. Pairwise accessory genomes using the Jaccard similarity index. Accessory-genome Jaccard similarity index (0 = no shared accessory genes, 1 = identical accessory gene content) and the number of accessory genes found exclusively in each pair and no other genome in the dataset, calculated from the Panaroo gene presence–absence matrix. Comparator pairs are the two other genome pairs that also received maximal (100) bootstrap support as sister taxa in the core-genome phylogeny

Genome pair	Bootstrap support[1]	Accessory-genome Jaccard index	Genes shared exclusively by this pair
HS11286 – DCP2	100	0.9956	436
KPNCDS21 (milk) – Kp-102 (Ganga river)	100	0.6972	56
257/KpS2A9 (lake) – IITR008 (river)	100	0.6956	21
HS11286 – IITR008 (river, ST147)	–	0.6216	0
HS11286 – Kp-102 (Ganga river)	–	0.6005	0

**Bootstrap support from the core-genome phylogeny (*Fig. 1*); – indicates the pair was not resolved as sister taxa in the tree.*

4. Discussion

This comparative genomic analysis identified an Indian river water *Klebsiella pneumoniae isolate*, DCP2, that is genomically indistinguishable at the clonal level from HS11286, an internationally significant, extensively characterized ST11 carbapenemase-producing clinical strain first described from a hospital in Shanghai, China. Four independent lines of evidence that is whole-genome ANI (99.99%), core-genome phylogenetic clustering (bootstrap = 100, near-zero branch length), identical plasmid replicon content at matching sequence identity, and a substantially elevated accessory-genome Jaccard similarity with 436 exclusively shared gene clusters, converge on the same conclusion: DCP2 and HS11286 represent the same clonal lineage, carrying the same high-risk resistance genotype (KPC-2, rmtB) on what appears to be the same mobile genetic background.

This finding is consistent with and extends a small but growing body of evidence that environmental water bodies can harbour genomically intact representatives of clinically significant, high-risk K. pneumoniae clones rather than only distantly related environmental relatives carrying isolated resistance genes [2, 5]. A second isolate in the present dataset, IITR008, independently confirmed this general pattern. This river-confluence isolate belongs to ST147, another internationally recognized high-risk lineage, and had previously been shown by others to cluster phylogenetically with clinical ST147 strains from India, Pakistan, and a traveler returning from India to the Czech Republic [2]. Taken together, the two river isolates characterized here indicate that at least two independently circulating high-risk clinical lineages (ST11 and ST147) have genomically confirmed environmental water counterparts within the same broad geographic setting.

The remaining environmental genomes in the dataset provide useful context for these two positive findings. All four water-source isolates DCP2, IITR008 and Kp-102 (river), and the lake isolate 257/KpS2A9, carried at least one Kleborate-detected acquired resistance gene (resistance score ≥ 1), whereas all four soil and milk isolates assigned to the *K. pneumoniae* species complex (KPNCDS21 from raw milk, and the rhizosphere, *K. variicola* and *K. quasipneumoniae* soil isolates) carried none (resistance score = 0). Separately, the five soil-associated genomes spanned five distinct taxa within the wider Klebsiella genus (*K. pneumoniae, K. variicola, K. quasipneumoniae, K. ornithinolytica and K. aerogenes*), indicating that soil in this setting harbors substantial genus-level diversity rather than a single dominant lineage. Taken together, these patterns are consistent with water, rather than soil, being the environmental compartment more commonly associated with elevated antimicrobial resistance burden in culturable *Enterobacteriaceae*, a pattern also reported for E. coli across water and soil sources in an unrelated agricultural watershed setting [14], though the small number of genomes per source category in the present dataset means this observation should be treated as descriptive rather than a formally tested association.

The present analysis also directly addresses a limitation explicitly identified in a recent large-scale global pan-genome and phylogenetic survey of *K. pneumoniae*, whose authors selected one representative genome per country and noted that this country-level sampling strategy could not resolve fine-grained relationships between individual isolates, recommending that future work integrate sequence typing and clonal lineage analysis [3]. By applying sequence typing, whole-genome ANI, and

plasmid replicon analysis within a single country across multiple environmental compartments, the present study provides exactly this finer-resolution view and identifies a specific, clonal-level clinical–environmental relationship that a broader, more sparsely sampled survey could not have detected.

From a One Health perspective, the shared plasmid replicon profile between DCP2 and HS11286 is a more informative observation than the presence of shared resistance genes alone [5]. Because KPC-2 and rmtB were found not only present in both genomes but embedded within an apparently identical plasmid backbone (three matching Inc replicon types at up to 100% identity), the most parsimonious interpretation is that DCP2 and HS11286 share a very recent common ancestor, whether through direct environmental contamination by a clinical-lineage strain, or persistence and circulation of this clone between clinical and environmental settings, rather than independent convergent acquisition of the same resistance genes by unrelated strains [5, 13].

This study has several limitations. First, the dataset comprises publicly deposited genome assemblies of heterogeneous contiguity (single-contig to > 100 contigs) and provenance; while BUSCO completeness was uniformly high, fragmentation is known to influence functional annotation counts in comparative genomic analyses, and resistance/virulence gene counts for the more fragmented assemblies should be interpreted with this in mind. Second, the sample size (n = 11, with mostly one genome per source category) precluded formal statistical hypothesis testing of gene enrichment by source; results are presented descriptively, supported by whole-genome ANI, phylogenetic bootstrap support, plasmid replicon identity, and pairwise accessory-genome Jaccard similarity as the primary quantitative measures of relatedness, consistent with standard practice in bacterial genomic epidemiology at this scale. Third, plasmid replicon typing was performed for HS11286 and DCP2 only; extending this analysis to the remaining environmental isolates and confirming physical linkage of resistance genes to the identified replicons (for example, by long-read sequencing or targeted contig binning) would further strengthen mechanistic inference and is recommended for future work. Fourth, the isolation source metadata for one genome in the dataset (DCP2) was found to differ between its associated peer-reviewed publication and its NCBI BioSample record; the BioSample record (river water, with georeferenced coordinates, submitted using the MIGS environmental sample metadata standard) was taken as authoritative for this analysis, and this discrepancy is noted here in the interest of transparency.

4. Conclusion

An Indian river water *Klebsiella pneumoniae* isolate was found to be genomically indistinguishable, at the level of whole-genome average nucleotide identity, core-genome phylogeny, and plasmid replicon content, from the international high-risk clinical clone HS11286. This provides direct genomic evidence that environmental water can harbour high-risk, carbapenemase-producing *K. pneumoniae* clones in a form essentially identical to their clinical counterparts, reinforcing the case for integrating environmental compartments into genomic surveillance of clinically significant antimicrobial resistance.

Declarations

Ethics, Consent to Participate, and Consent to Publish

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable. This study used only publicly available genome sequence data retrieved from the National Center for Biotechnology Information (NCBI). No human participants, animal subjects, or identifiable personal data were involved, and no new biological samples were collected.

Data Availability

All genome assemblies analyzed in this study are publicly available from NCBI RefSeq under the accessions listed in **Table 1**.

Author Contributions

N.A. performed all Bioinformatics and statistical analyses, curated the data, prepared the figures, and wrote the manuscript. **S.S.D.** designed the study and contributed significantly to data validation and manuscript revision.

Transparency declarations

This work received no external funding and was carried out as independent academic research by the authors.

None of the authors has any conflicts of interest to declare in relation to this work.

References

1. Wyres KL, Holt KE. Klebsiella pneumoniae Population Genomics and Antimicrobial-Resistant Clones. Trends Microbiol. 2016;24(12):944–56.
2. Singh D, et al. Comparative genomics of an extensively drug resistant strain Klebsiella pneumoniae IITR008 with international high-risk clonal lineage ST147 isolated from river water. Antonie Van Leeuwenhoek. 2024;117(1):57.
3. Niyazi HA, et al. Pan-genome analysis and phylogenetic characterization of Klebsiella pneumoniae from global isolates. Future Microbiol. 2026;21(8):701–12.
4. Liu P, et al. Complete genome sequence of Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain isolated from human sputum. J Bacteriol. 2012;194(7):1841–2.
5. Runcharoen C, et al. Whole genome sequencing reveals high-resolution epidemiological links between clinical and environmental Klebsiella pneumoniae. Genome Med. 2017;9(1):6.
6. Sanjay A, et al. Unveiling Genetic Insights of Klebsiella Pneumoniae with Whole Genome Sequence Analysis. Acta Sci Microbiol. 2024;ISSN(11):2581–3226.
7. Lam MMC, et al. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. Nat Commun. 2021;12(1):4188.
8. Tonkin-Hill G, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21(1):180.
9. Minh BQ, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530–4.
10. Jain C, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.
11. Alcock BP, et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. Nucleic Acids Res. 2023;51(D1):D690–9.
12. Liu B, et al. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 2019;47(D1):D687–92.
13. Carattoli A, Hasman H. *PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS).* Methods Mol Biol, 2020. 2075: pp. 285–294.
14. Mukherjee M, et al. Escherichia coli antimicrobial resistance variability in water runoff and soil from a remnant native prairie, an improved pasture, and a cultivated agricultural watershed. Water. 2020;12(5):1251.

Figures

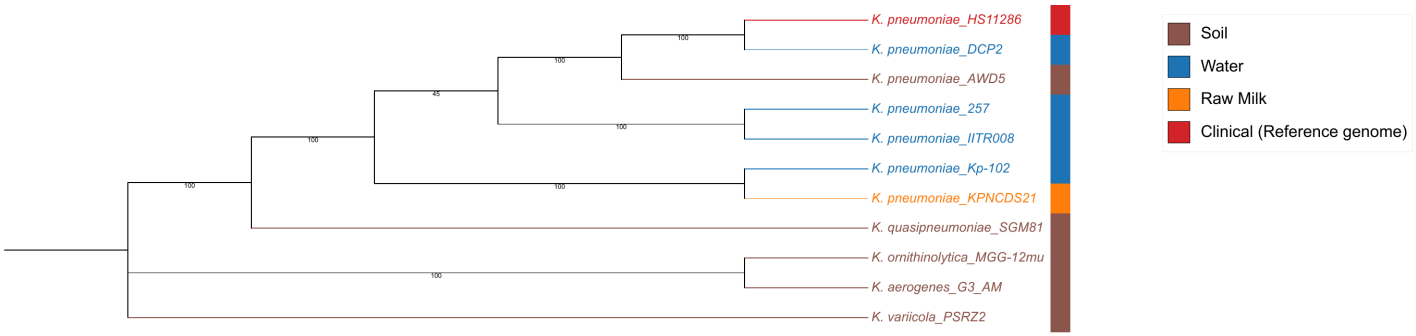


Figure 1

Core-genome maximum-likelihood phylogeny of the eleven-genome dataset (IQ-TREE) from the Panaroo core-gene alignment. Coloured bars indicate isolation source (soil, water, raw milk, reference). Node support (ultrafast bootstrap approximation, percentage values shown at internal nodes) was assessed in IQ-TREE. HS11286 and DCP2 form a fully supported (bootstrap = 100) sister pair separated by a near-zero branch length (0.00003 substitutions/site), the shortest branch separation in the tree.

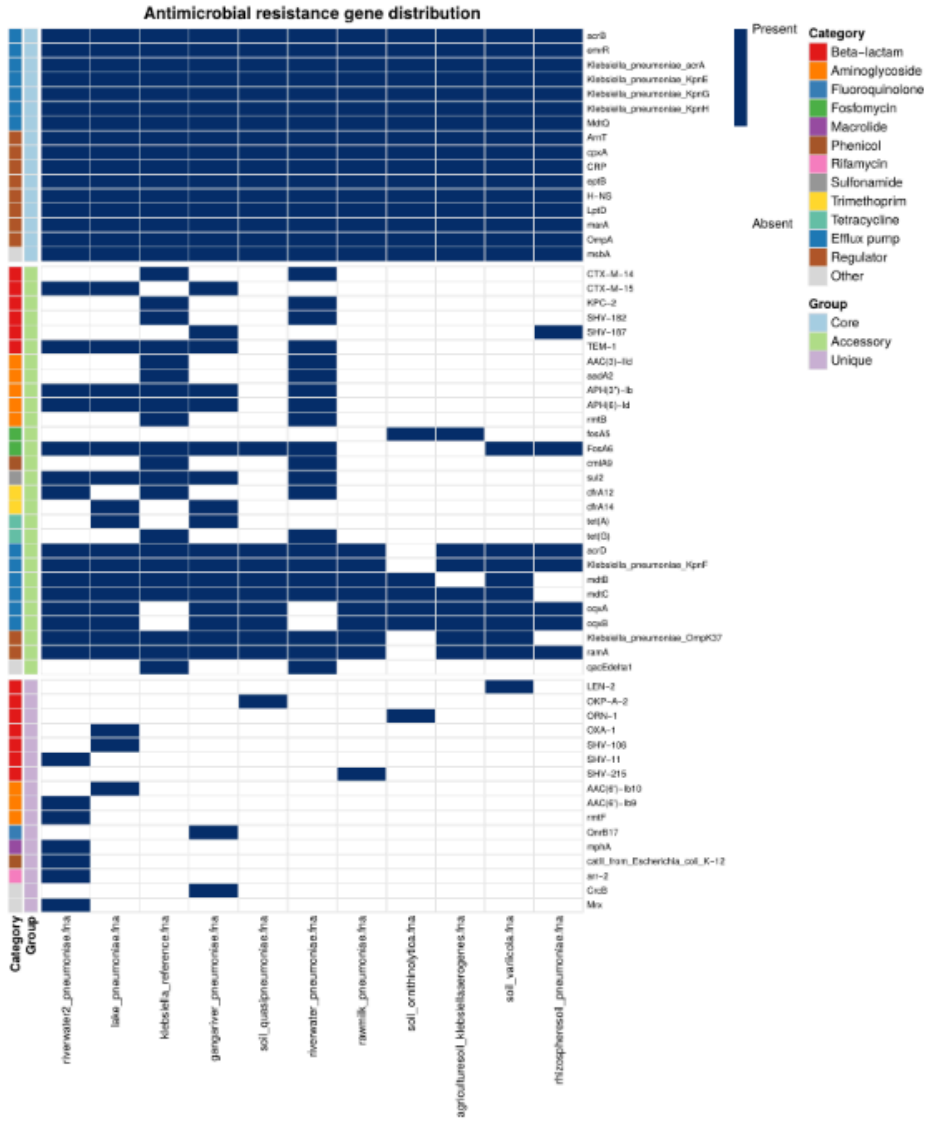


Figure 2

Antimicrobial resistance gene presence–absence across the eleven-genome dataset dataset, identified against the Comprehensive Antibiotic Resistance Database (CARD). Genes are ordered by conservation category (Core, present in all 11 genomes; Accessory, present in 2–10 genomes; Unique, present in exactly one genome) and by resistance class; genomes are shown in a fixed order and were not subjected to hierarchical clustering.

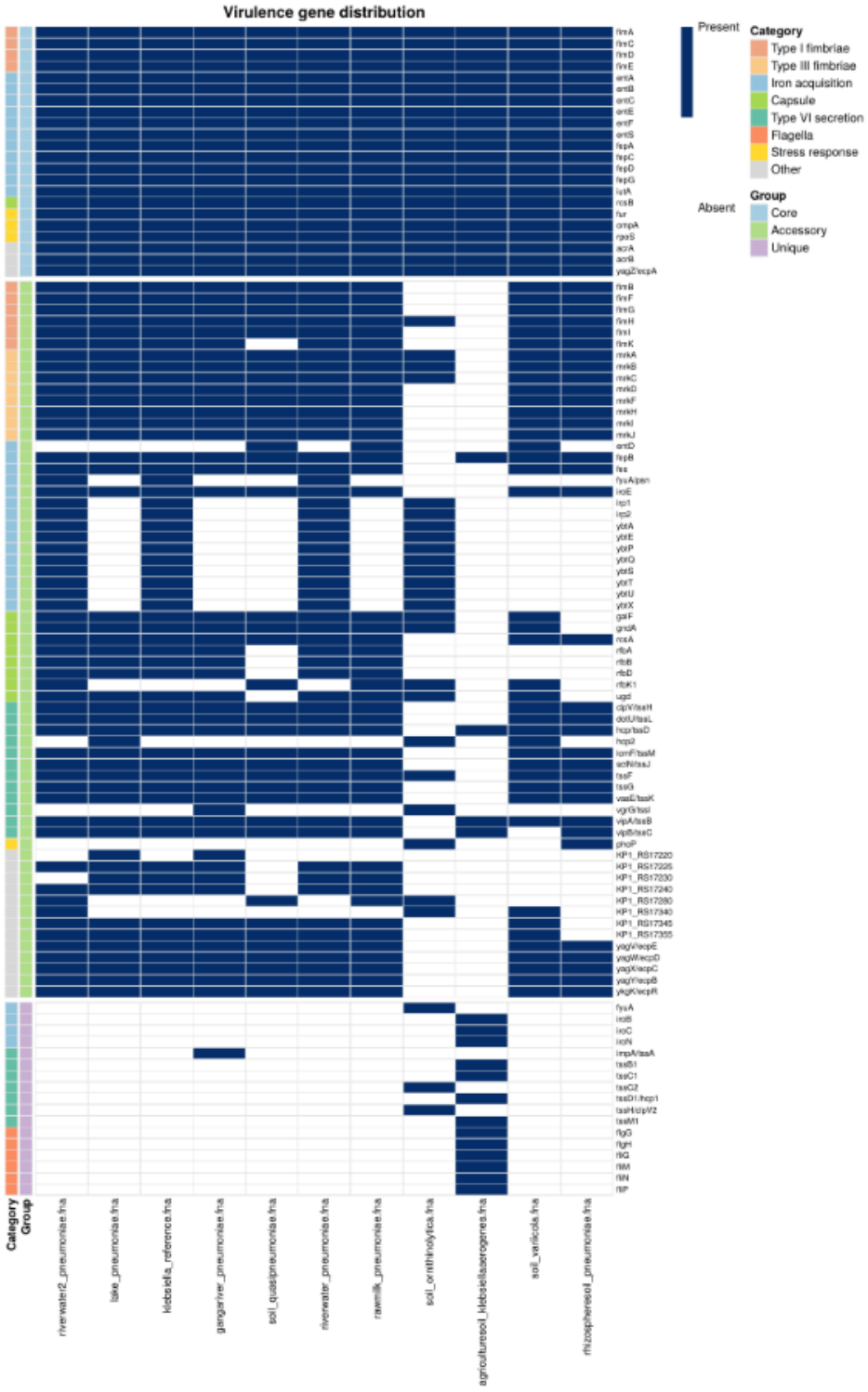


Figure 3

Virulence-associated gene presence–absence across the eleven-genome dataset, identified against the Virulence Factor Database (VFDB). Genes are ordered by conservation category (Core, present in all 11 genomes; Accessory, present in 2–10

genomes; Unique, present in exactly one genome) and by functional category (e.g. fimbrial, iron acquisition, capsule, type VI secretion); genomes are shown in a fixed order and were not subjected to hierarchical clustering.