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.

A project-selection audit of public urinary-source Escherichia coli genomes from China distinguishes purpose-enriched NDM and mcr signals from CTX-M recurrence across projects

Botao Yu
Ningbo Medical Center Lihuili Hospital
Chunling Wang
Ningbo Medical Center Lihuili Hospital
Zhen Wang
Ningbo Medical Center Lihuili Hospital
Min Yin
yinminmed@outlook.com

Ningbo Medical Center Lihuili Hospital

Research Article

Keywords: Escherichia coli, urinary tract infection, antimicrobial resistance, public genomes, BioProject selection, genomic epidemiology, plasmids, integrons

Posted Date: October 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10841418/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

Public bacterial genome archives combine surveillance collections, resistance-focused studies, selected clinical subsets, and individual genome reports. The reasons for sequencing can therefore affect pooled resistance proportions. This study traced project-level selection mechanisms in public urinary-source *Escherichia coli* genomes from China and assessed clonal and mobile resistance signals.

Results

The cohort contained 512 genomes from 29 BioProjects. Original publications were identified for 19 projects that contributed 362 genomes. BioProject or case records supported the remaining 10 projects and 150 genomes. Eighteen resistance-targeted or enriched projects supplied 350 genomes (68.4%). All 144 NDM-positive genomes and 142 of 145 mcr-positive genomes arose from this stratum. CTX-M occurred in 25 of 29 projects. It was present in 44/92 genomes (47.8%) from a syndrome-defined clinical collection and 32/55 (58.2%) from two representative clinical subsets. Its pooled proportion varied by no more than 4.2 percentage points in leave-one-project-out analyses. Major Achtman lineages showed distinct resistance profiles. ST410 and ST167 were NDM-rich. ST131 and ST1193 mainly carried fluoroquinolone resistance-associated and CTX-M genotypes. A recombination-filtered ST167 phylogeny recovered several related NDM-positive sublineages. Local sequence contexts associated NDM mainly with IncX3-like regions and mcr mainly with IncI2- and IncX4-like regions. Recurrent integron cassette modules occurred across multiple projects and lineages.

Conclusions

The project-selection audit separated concentrated NDM and mcr signals from CTX-M recurrence across projects. Source-study audits should accompany accession-level analysis of heterogeneous public genome collections. Together, these analyses estimate archive composition while retaining clonal and mobile-element context.

Background

Urinary tract infections are among the most common bacterial infections, and *Escherichia coli* is their predominant cause [1]. Antimicrobial resistance complicates empirical treatment and has increased the use of whole-genome sequencing for surveillance and outbreak investigation. Earlier multicentre work in China documented extensive resistance and substantial clonal diversity among urinary *E. coli* isolates [2]. Public archives now extend these collections across years, regions, lineages, and resistance mechanisms.

Public bacterial genome archives are convenience collections assembled under heterogeneous selection criteria [3]. A BioProject may contain a consecutive clinical collection, an isolate subset chosen to represent phenotypes, a resistance-gene-positive cohort, or a single unusual strain. These choices can strongly

influence pooled archive proportions. Project size alone does not show whether isolates were selected by syndrome, phenotype, resistance gene, or clone. Selection and reporting biases are established concerns in antimicrobial-resistance surveillance [4].

Population structure introduces a second layer of dependence. The Achtman multilocus sequence typing scheme defines widely used *E. coli* lineages [5]. ST131 and ST1193 are established extraintestinal lineages associated with fluoroquinolone resistance and extended-spectrum beta-lactamases [6, 7]. Carbapenem-resistant collections from China also contain ST410, ST167, ST405, and other lineages [8]. Global analyses identify ST167 and ST410 as recurrent hosts of *bla*NDM [9]. Mobile determinants cross these lineage boundaries. The emergence of *mcr-1* in China showed how plasmids disseminate colistin resistance among bacterial backgrounds [10, 11]. Integrons assemble and redistribute resistance cassettes [12].

Auditing project selection identifies signals concentrated by study design. Lineage-specific phylogeny, plasmid-associated local context, and cassette organization then resolve clonal and mobile patterns.

An accession-level cohort was assembled from 512 public urinary-source *E. coli* genomes from China. The selection mechanism of every contributing BioProject was audited through its original publication and public project record. Genomic results were stratified by selection design and subjected to project-level sensitivity analyses. Subsequent analyses characterized major Achtman lineages, reconstructed a recombination-filtered ST167 phylogeny, compared NDM and mcr local contexts, and identified recurrent integron modules. The target estimand was archive composition by project design.

Methods

Study design and cohort definition

This was an accession-level computational analysis of public bacterial genomes. Public read and assembly records were screened for *E. coli*, a China-associated geographic field, urinary-source evidence, and whole-genome data. Candidate records were reconciled across run, BioSample, BioProject, and assembly accessions. Each retained genome received one cohort identifier. Structured specimen-source fields supported 506 records. Six records required manual review of the associated BioSample, run, or BioProject context. The accession ledger retained country, province or city, collection year, host, data source, and inclusion evidence (Additional file 2). The frozen cohort contained 512 genomes from 29 BioProjects.

The accession-level genome was the primary unit of analysis. Potentially linked or duplicated records were retained because public metadata could not reliably distinguish replicate sequencing, resubmission, and repeated isolation. BioProject was the clustering unit for archive-structure analyses. The study included all public records that met the frozen eligibility criteria, so no prospective sample-size calculation was performed.

BioProject selection-mechanism audit

Public-record searching began on 2 July 2026 during cohort construction. After the accession cohort and genomic outcome calls were frozen, each of the 29 retained BioProject accessions was checked in Europe PMC and the NCBI public records. The review covered the project title and description, linked publications, study methods, and stated isolate inclusion criteria. An original publication was the preferred source. When no linked paper could be identified, the BioProject record or an accession-linked case report served as the source. The audit ledger retained the source, quoted selection evidence, expected archive effect, and uncertainty (Additional file 3).

Projects were assigned to one of four prespecified descriptive classes. A syndrome-defined clinical collection enrolled patients or isolates according to a clinical syndrome without an explicit resistance-gene entry criterion. A representative clinical subset selected genomes from a broader clinical collection to span molecular types, phenotypes, or comparison groups. A resistance-targeted or enriched project required, screened for, or was organized around a resistance phenotype, resistance gene, or multidrug-resistant organism. A lineage, case, or genome-report collection consisted of lineage-focused sets, paired isolates, technical demonstrations, or individual genomes without a population sampling denominator. The most explicit documented entry criterion determined the class. Genotype proportions within these strata were treated as measures of archive composition.

The audit linked 19 BioProjects to original publications. These included mcr- or NDM-targeted collections [13–18], carbapenem-resistant or ESBL-producing cohorts [19–24], broader clinical subsets [25, 26], and lineage or technical-case studies [6, 27–30]. Ten projects had no accession-resolved original publication in the search and were classified from their BioProject or case records. Full source links and selection statements are provided in Additional file 3.

Read processing, assembly, and quality control

Public reads were retrieved with SRA Toolkit 3.4.1 and preprocessed with fastp 1.3.6 [31]. SeqKit 2.13.0 was used for FASTA and FASTQ manipulation [32]. Reads were assembled with Shovill 1.4.2 using an expected genome size of 5.1 Mb. When suitable reads were unavailable, the corresponding NCBI assembly was retained. The final cohort included 427 read-based reassemblies and 85 retained public assemblies.

Assembly statistics were generated with QUAST 5.3.0 [33] and SeqKit. Assemblies outside 4.3 to 6.5 Mb failed the cohort gate. Warning flags were applied to assemblies outside 4.5 to 6.2 Mb, with more than 700 contigs, or with an N50 below 20 kb. Taxonomic quality control used Kraken2 2.17.1 and the standard full database dated 26 February 2026 [34]. A strict pass required an *E. coli* fraction of at least 0.80 and a non-*E. coli* fraction no greater than 0.20. All 512 accession-level genomes remained in the primary cohort. The 506 strict passes formed a sensitivity set.

Typing and resistance-genotype definitions

Achtman sequence types were assigned with mlst 2.32.2 and the ecoli_achtman_4 scheme. This nomenclature was used for all lineage comparisons. ClermonTyping 24.02 assigned phylogroups [35].

AMRFinderPlus 4.2.7 with database version 2026-03-24.1 detected acquired resistance genes and resistance-associated mutations [36]. Analyses used the *Escherichia* organism setting and the --plus option. Plasmid replicons were screened with ABRicate 1.4.0 against the PlasmidFinder database frozen on 1 May 2026.

CTX-M, NDM, and mcr flags were assigned from exact family-normalized AMRFinderPlus results. A fluoroquinolone resistance-associated genotype required a quinolone-target mutation, a qnr marker, or aac(6')-Ib-cr. The MDR-like definition covered ten resistance-marker categories: extended-spectrum beta-lactamase-like, carbapenemase, mcr-associated colistin resistance, fluoroquinolone, aminoglycoside, fosfomycin, sulfonamide or trimethoprim, tetracycline, macrolide, and phenicol. Genomes positive in at least three categories were classified as MDR-like. These categories served as comparative genomic definitions. Harmonized antimicrobial-susceptibility phenotypes were unavailable.

Mobile-element and plasmid analyses

Assemblies were annotated with Prokka 1.15.6 [37]. Mobile-element prediction used geNomad 1.12.0 with database v1.9 [38], ISEScan 1.7.3 [39], IntegronFinder 2.0.6 [40], and MOB-recon from MOB-suite 3.1.9 [41]. IntegronFinder calls included complete integrons and CALINs. MOB-recon identified predicted conjugative plasmids. Replicon, insertion-sequence, integron, and plasmid-bin evidence was retained at sample and contig levels.

Pangenome analysis and broad phylogeny

Panaroo 1.8.0 was run on 512 standardized Prokka GFF files with strict cleaning, MAFFT core alignment, and a 0.95 core threshold [42]. The filtered core alignment was 3,042,123 bp. Sites containing A, C, G, or T were retained, and other characters were recoded as N. Sites with more than 5% missing data were removed. The resulting variable-site alignment contained 329,173 positions.

IQ-TREE 2.4.0 inferred an unrooted maximum-likelihood tree under GTR + F+ASC+R10 with 1,000 ultrafast-bootstrap replicates, 1,000 SH-aLRT replicates, and BNNI optimization [43]. Sequence type, BioProject, resistance genotype, selection class, and quality-control labels were excluded from alignment filtering and topology optimization. These annotations were added after inference, and the broad tree summarized population structure and genotype distribution.

ST167 recombination-filtered phylogeny and duplicate-linked audit

The 28 Achtman ST167 assemblies were called against the study-selected reference UPEC_CHN_0202 with Snippy 4.0.2. Minimum base quality was 10, minimum depth was 10, and minimum allele fraction was 0.9. Gubbins 3.4.3 was applied to the core alignment [44], leaving 1,742 polymorphic sites. IQ-TREE 2.4.0 selected TVMe + ASC and generated 1,000 ultrafast-bootstrap and 1,000 SH-aLRT replicates. Pairwise distances were computed with snp-dists 1.2.0.

Pairs separated by no more than five filtered polymorphisms were reviewed using accession, BioProject, location, year, strain labels, and Mash similarity. The 28 accession-level genomes remained the primary set. Sensitivity analyses used the strict taxonomic quality-control subset and a second set with one representative removed from each of the two strongest probable duplicate-linked pairs. Pairwise distributions were summarized descriptively because pairs share genomes. Further criteria are given in the Supplementary Methods (Additional file 1).

NDM and mcr local-context analysis

Target contigs were extracted from the evidence table and reanalyzed with AMRFinderPlus 4.2.7. Prokka annotated features within 20 kb on either side of the target. ISEScan calls within 10 kb were joined to same-contig replicon, integron, and geNomad evidence. BLAST + 2.17.0 compared local contigs with 62 cached NCBI plasmid references. Hits were ranked by bit score and then query coverage. A strict reference match required at least 99% identity, 80% query coverage, and 80% reference coverage. Representatives retained distinct target variants, replicons, reference matches, sequence types, projects, and insertion-sequence arrangements. Local-context conclusions were restricted to the assembled target contigs.

Integron cassette reconstruction

IntegronFinder sample-level calls were joined to Prokka coding-sequence coordinates. AMRFinderPlus supplied standardized resistance-gene names for cassette proteins. Features in complete integrons were ordered outward from intI. CALIN signatures were canonicalized in both directions because they lack an integrase orientation anchor. A recurrent module required at least five genomes and two BioProjects. No single project could contribute more than 80%. Each module also required at least two component genes, two cassettes, a mean mapped fraction of at least 0.80, and no hypothetical component in its signature. CALIN orientation was left unspecified.

Statistical and sensitivity analyses

Primary proportions used all 512 genomes. Strict-quality sensitivity analyses used 506. Project-equalized summaries were unweighted means of the 29 BioProject-specific proportions. For each outcome, dominant-project exclusion removed the single BioProject that produced the largest absolute change from the pooled proportion. Leave-one-BioProject-out estimates were calculated for every project (Additional file 3).

Selection-class summaries used accession-level counts and denominators within the four documented project-design classes. The selection-class origin of every marker-positive genome was also counted. These post hoc summaries quantify archive composition within the documented classes. Residual selection remains within the syndrome-defined collection and representative subsets. Two-sided Fisher exact tests were exploratory and are reported in Additional file 8. Additional methods and software parameters are provided in Additional files 1 and 7.

Results

Most archived genomes came from resistance-targeted projects

The final cohort contained 512 urinary-source *E. coli* genomes from 29 BioProjects (Fig. 1; Table 1; Additional file 2). Collection year was available for 497 genomes and ranged from 2006 to 2025. A province or city was recorded for 470 genomes, and human host metadata were explicit for 507. Raw public reads were reassembled for 427 genomes. An NCBI assembly was retained for 85. The Kraken2 screen classified 506 genomes as strict taxonomic quality-control passes.

Table 1
Project selection classes and resistance-genotype composition

Project selection class	BioProjects	Genomes	CTX-M, n (%)	NDM, n (%)	mcr, n (%)	FQ genotype, n (%)	MDR-like, n (%)
Syndrome-defined clinical collection	1	92	44 (47.8)	0 (0.0)	1 (1.1)	86 (93.5)	59 (64.1)
Representative clinical subset	2	55	32 (58.2)	0 (0.0)	0 (0.0)	53 (96.4)	36 (65.5)
Resistance-targeted or enriched	18	350	256 (73.1)	144 (41.1)	142 (40.6)	338 (96.6)	335 (95.7)
Lineage, case, or genome-report collection	8	15	12 (80.0)	0 (0.0)	2 (13.3)	15 (100.0)	12 (80.0)
Total public archive	29	512	344 (67.2)	144 (28.1)	145 (28.3)	492 (96.1)	442 (86.3)

Percentages use genomes within each row as the denominator. The four classes describe documented project selection mechanisms. Values are archive-composition measures and are not corrected prevalence estimates. FQ, fluoroquinolone resistance-associated genotype.

Original-publication evidence was available for 17 projects and 353 genomes. Two further projects with nine genomes were supported by an original publication together with the BioProject record. One genome was supported by a BioProject record and case report. The remaining nine projects and 149 genomes relied on BioProject records alone. Thus, an original publication was identified for 19/29 projects and 362/512 genomes. The source trail for every project is provided in Additional file 3.

Project purpose was highly uneven (Fig. 1A). Eighteen resistance-targeted or enriched projects contributed 350 genomes (68.4%). One syndrome-defined clinical collection contributed 92 genomes, and two representative clinical subsets contributed 55. Eight lineage, case, or genome-report projects contributed the remaining 15 genomes. The five largest BioProjects accounted for 379/512 genomes (74.0%) (Fig. 1B).

The pooled archive contained CTX-M in 344/512 genomes (67.2%), NDM in 144/512 (28.1%), and mcr in 145/512 (28.3%). A fluoroquinolone resistance-associated genotype occurred in 492/512 (96.1%), and

442/512 (86.3%) met the MDR-like genotype definition. These proportions differed by project selection class (Fig. 1C; Table 1). The resistance-targeted stratum contained 256/350 CTX-M-positive genomes (73.1%), all 144 NDM-positive genomes, 142/350 mcr-positive genomes (40.6%), and 335/350 MDR-like genomes (95.7%).

The less explicitly resistance-enriched clinical strata comprised 147 genomes. CTX-M was present in 44/92 genomes (47.8%) from the syndrome-defined collection and 32/55 (58.2%) from the representative subsets. Together, these strata contained 76/147 CTX-M-positive genomes (51.7%), no NDM-positive genomes, and one mcr-positive genome (0.7%). Their fluoroquinolone resistance-associated and MDR-like genotype proportions were 139/147 (94.6%) and 95/147 (64.6%), respectively.

NDM and mcr signals were concentrated in purpose-enriched projects

The origin of marker-positive genomes differed sharply among genotype classes (Fig. 2A). CTX-M occurred in 25/29 BioProjects. Resistance-targeted or enriched projects supplied 256/344 CTX-M-positive genomes (74.4%). They supplied all 144 NDM-positive genomes and 142/145 mcr-positive genomes (97.9%). The syndrome-defined collection and representative subsets contributed 76/344 CTX-M-positive genomes (22.1%).

Project-level sensitivity results followed the same pattern (Fig. 2B; Additional file 3). The CTX-M proportion was 67.2% in the pooled cohort and strict-quality subset and 68.5% after project equalization. Removing the most influential project yielded 300/420 CTX-M-positive genomes (71.4%). The maximum leave-one-project-out change was 4.2 percentage points.

NDM was present in 28.1% of the pooled archive and 28.5% of the strict-quality subset. The project-equalized mean was 37.3%, reflecting high NDM proportions in several small, targeted projects. Excluding PRJNA931432 reduced NDM to 80/448 genomes (17.9%). The pooled mcr proportion was 28.3%, and the strict-quality value was 27.9%. Project equalization reduced it to 18.1%. Excluding the mcr-1-positive collection PRJNA604918 reduced it to 25/392 genomes (6.4%).

Major Achtman lineages carried distinct resistance profiles

Achtman sequence types were callable for 502/512 genomes (98.0%). The most frequent were ST410 (67 genomes), ST131 (60), ST1193 (41), ST167 (28), ST69 (22), ST648 (20), ST38 (19), and ST405 (16) (Table 2; Additional file 4). Panaroo identified 20,262 gene clusters, including 3,129 core genes at the prespecified threshold.

Table 2
Resistance-genotype profiles of major Achtman lineages

Achtman lineage	Genomes	BioProjects	Dominant phylogroup	CTX-M, n (%)	NDM, n (%)	mcr, n (%)	FQ genotype, n (%)	MDR-like, n (%)
ST410	67	13	C	51 (76.1)	56 (83.6)	3 (4.5)	67 (100.0)	67 (100.0)
ST167	28	10	A	24 (85.7)	21 (75.0)	12 (42.9)	28 (100.0)	28 (100.0)
ST131	60	14	B2	43 (71.7)	2 (3.3)	2 (3.3)	60 (100.0)	47 (78.3)
ST1193	41	4	B2	22 (53.7)	0 (0.0)	1 (2.4)	41 (100.0)	22 (53.7)
ST405	16	10	D	14 (87.5)	8 (50.0)	2 (12.5)	15 (93.8)	15 (93.8)
ST648	20	7	F	14 (70.0)	3 (15.0)	8 (40.0)	20 (100.0)	17 (85.0)
ST69	22	7	D	8 (36.4)	1 (4.5)	4 (18.2)	19 (86.4)	16 (72.7)
ST38	19	6	D	16 (84.2)	2 (10.5)	2 (10.5)	18 (94.7)	18 (94.7)

Percentages use the number of genomes in each lineage as the denominator. FQ, fluoroquinolone resistance-associated genotype.

Aggregate ST410 contained 56/67 NDM-positive genomes (83.6%), 51/67 CTX-M-positive genomes (76.1%), and 3/67 mcr-positive genomes (4.5%) (Fig. 2C). In ST167, 21/28 genomes were NDM positive (75.0%), 24/28 carried CTX-M (85.7%), and 12/28 carried mcr (42.9%). Every genome in both lineages had a fluoroquinolone resistance-associated genotype and met the MDR-like genotype definition.

ST131 and ST1193 had different profiles. Both were uniformly positive for the fluoroquinolone resistance-associated genotype. NDM occurred in 2/60 ST131 genomes and no ST1193 genome. CTX-M occurred in 43/60 ST131 genomes (71.7%) and 22/41 ST1193 genomes (53.7%). ST405 contained 8/16 NDM-positive genomes, while mcr occurred in 8/20 ST648 genomes. Project structure and lineage were related. PRJNA931432 contributed 47/67 ST410 genomes, and excluding it reduced the archived ST410 proportion from 67/512 (13.1%) to 20/448 (4.5%).

Broad phylogeny recovered major lineage clusters

The maximum-likelihood tree was inferred from 329,173 variable sites in the Panaroo core alignment (Fig. 3; Additional files 4 and 9). Major ST410, ST131, and ST1193 clusters were evident. CTX-M occurred across multiple branches. NDM formed dense lineage-project concentrations, while mcr was distributed among lineages and was especially common in the large PRJNA604918 block.

IQ-TREE reported composition chi-square warnings for 349/512 sequences in the broad alignment. Overall missingness was 0.29%, and the maximum per-genome value was 2.26%.

ST167 contained several related NDM-positive sublineages

The ST167-specific analysis included 28 genomes from 10 BioProjects. Reference-based calling followed by Gubbins recombination filtering retained 1,742 polymorphic sites. The maximum-likelihood tree resolved several NDM-positive branches (Fig. 4A; Additional files 4 and 10). Across all ST167 pairs, the median filtered distance was 161 SNPs (interquartile range, 85 to 236.8). It was 89.5 for NDM-positive pairs (interquartile range, 69 to 112), 210 for NDM-discordant pairs (179.5 to 248.5), and 192 for NDM-negative pairs (153 to 314) (Fig. 4B).

Four pairs separated by zero or one filtered SNP were reviewed. One cross-project pair was judged a probable duplicate or resubmission. A second pair was a probable linked isolate or resequencing event. Metadata were insufficient to remove the other two pairs. The 28-genome analysis remained primary. The median for NDM-positive pairs was 89.5 SNPs in the strict taxonomic subset and 90 after one representative from each of the two strongest duplicate-linked pairs was removed. The corresponding NDM-discordant medians were 243 and 210 SNPs (Supplementary Results in Additional file 1).

NDM and mcr occurred in distinct plasmid-associated local contexts

Target-contig analysis recovered 359 unique contigs from 217 genomes. AMRFinderPlus confirmed 138 exact target-gene records with sufficient local context from 132 genomes, comprising 38 NDM and 100 mcr records (Additional file 5). The NDM contexts included NDM-5 (28), NDM-1 (8), NDM-4 (1), and NDM-9 (1). An insertion sequence was detected within 10 kb in 37/38 records. IncX3 occurred on the same contig in 22/38. Ten records met the strict plasmid-reference criteria, and all were most similar to the IncX3 plasmid JN247852.1 (Fig. 5A).

The 100 exact mcr contexts comprised mcr-1.1 (92), mcr-1 (6), mcr-1.26 (1), and mcr-1.8 (1). IncI2 occurred on the same contig in 69 records and IncX4 in 24. Twenty records met the strict reference criteria. IncI2 representatives matched regions of MH202956.1, CP042471.2, CP031550.1, and KY120366.1 (Fig. 5B). One strict IncX4 representative matched MH061196.1 (Fig. 5C). Eighty-two of the 100 exact mcr local-context records came from PRJNA604918.

Recurrent integron cassette modules spanned projects and lineages

IntegronFinder identified an integron or CALIN in 377/512 genomes (73.6%), and MOB-suite predicted at least one conjugative plasmid in 393/512 (76.8%). The integron-positive genomes contained 634 integron or CALIN instances and 1,407 cassette proteins. Coordinate mapping resolved five recurrent cassette organizations (Supplementary Fig. 1; Additional file 6).

The dfrA17-aadA5 module occurred in 108 genomes from 19 BioProjects. The dfrA12-aadA2 module occurred in 97 genomes from 14 BioProjects, including 46 ST410 and 51 non-ST410 genomes. Fifty-three of

these 97 occurrences were outside PRJNA931432. The longer dfrA12-aadA2-cmlA1-aadA1-qacL module occurred in 21 genomes from nine BioProjects. Two less frequent modules were dominated by CALIN calls. The aac(6')-Ib-cr-blaOXA-catB3-arr-3 module occurred in 26 genomes from 10 projects. The aadA22-lnu(F) module occurred in seven genomes from six projects.

Discussion

Project selection explained the strongest contrast in this archive. More than two-thirds of the genomes came from studies that selected resistance phenotypes, resistance genes, or multidrug-resistant organisms. This stratum contained every NDM-positive genome and all but three mcr-positive genomes. CTX-M occurred in 25 BioProjects, including the syndrome-defined collection and both representative subsets. Its pooled proportion varied by no more than 4.2 percentage points in leave-one-project-out analyses. The NDM and mcr proportions primarily describe project enrichment, whereas CTX-M recurred across different project designs.

Project purpose added information that project size alone could not supply. The largest project was restricted to *mcr-1*-positive *E. coli* [13]. The dominant NDM project arose from a carbapenem-resistant collection assembled to study ST410 [19]. Other studies selected NDM-positive isolates [14, 16], carbapenemase-producing *E. coli* or carbapenem-resistant lineages [21, 22], ESBL-producing *E. coli* [20, 23, 24], or unusual multidrug-resistant genotypes [15]. These collections support clone and mobile-element comparisons. Their pooled counts have no population prevalence denominator.

The less explicitly resistance-enriched collections sharpened the interpretation of CTX-M. The syndrome-defined collection enrolled patients with type 2 diabetes. The two representative subsets selected genomes from larger clinical collections to compare molecular types or phenotypic groups [25, 26]. CTX-M occurred in 76/147 genomes (51.7%) across these collections. Its recurrence in both project types and stability under project-level sensitivity analyses distinguish CTX-M from the concentrated NDM and mcr signals.

Lineage stratification identified distinct clonal concentrations. ST131 and ST1193 retained their recognized fluoroquinolone- and CTX-M-associated profiles, consistent with their established international roles [6, 7]. Aggregate ST410 and ST167 were NDM-rich, as reported in previous genomic studies [9, 45, 46]. One carbapenem-resistant BioProject supplied most of the ST410 signal [19]. ST167 occurred in 10 projects, and its recombination-filtered tree recovered several related NDM-positive branches. The distance pattern persisted after removal of the strongest duplicate-linked pairs. These results localize the main clonal resistance signals within the archive.

The broad tree supported lineage-level interpretation. Composition warnings affected 349 sequences despite low missingness and independent taxonomic quality control. The topology displayed broad lineage clusters and resistance distributions. The recombination-filtered ST167 tree supplied higher resolution for that lineage.

Local contexts matched recognized mobile platforms. IncX3 plasmids are important vehicles for *bla*NDM in Enterobacterales in China [47]. IncX3/JN247852.1-like regions occurred with several sequence types and

NDM variants. IncI2 and IncX4 are established *mcr-1* platforms [11, 48], and they were the main same-contig replicon associations in this cohort. These contig-level matches identify related local backbones. Complete plasmid structure and transfer relationships require long-read closure.

Integron modules supplied another cross-project signal. The dfrA17-aadA5 and dfrA12-aadA2 organizations are common class 1 integron-associated arrangements [12, 49]. The three most frequent modules occurred across 9 to 19 BioProjects and multiple sequence types. More than half of the dfrA12-aadA2 occurrences were outside the dominant NDM project. The two less frequent modules were dominated by CALIN calls, which lack an integrase orientation anchor. These CALIN results define cassette composition, with orientation left unresolved.

The cohort had uneven geographic and temporal coverage and no probability sampling. Source tracing identified original publications for 19 of 29 projects. The remaining classifications relied on BioProject or case records with less detail. The four design classes summarize project entry criteria but leave some within-project selection steps unresolved. Genomic resistance markers lacked harmonized antimicrobial-susceptibility phenotypes, and public metadata left some duplicate links unresolved. Composition heterogeneity limits the broad tree to lineage-level interpretation. Short-read contigs limit plasmid analysis to local context.

Public-genome studies should audit project purpose alongside project size. In this archive, the source-study audit separated purpose-enriched NDM and mcr signals from CTX-M recurrence across project designs. Accession-level genomics then resolved NDM-rich clonal groups, IncX3-, IncI2-, and IncX4-associated local contexts, and recurrent integron modules. Together, these analyses define both the composition and biological structure of the archive.

Conclusions

Sequencing-project selection structured the resistance signals in this public archive. Resistance-targeted studies supplied all NDM-positive genomes and 142 of 145 mcr-positive genomes. CTX-M occurred across 25 projects and remained common in less explicitly enriched clinical collections. Linking source-study audit to accession-level genomics distinguishes purpose-enriched signals from those recurring across project designs. It also resolves clonal lineages, local mobile-element contexts, and recurrent cassette modules.

Abbreviations

AMR
antimicrobial resistance
CALIN
cluster of attC sites lacking a nearby integrase
CRE
carbapenem-resistant Enterobacterales
ESBL
extended-spectrum beta-lactamase

FQ
fluoroquinolone
MDR
multidrug resistance
MLST
multilocus sequence typing
NDM
New Delhi metallo-beta-lactamase
SNP
single-nucleotide polymorphism
UTI
urinary tract infection.

Declarations

Ethics approval and consent to participate

Not applicable. The study used deidentified, publicly available bacterial genomic records and did not recruit participants, perform interventions, or access identifiable human information.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Medical Science and Technology Project of Zhejiang Province (Grant No. 2024638739), awarded to B.Y. The funding body had no role in the design of the study, data collection, analysis, interpretation, manuscript preparation, or the decision to submit the work for publication.

Author Contribution

B.Y. contributed to conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, and writing the original draft. C.W. contributed to methodology, validation, and review and editing of the manuscript. Z.W. contributed to data curation, validation, and review and editing of the manuscript. M.Y. contributed to conceptualization, funding oversight, project administration, supervision, and review and editing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Data Availability

The datasets supporting the conclusions of this article are included within the article and its additional files. Accession-level NCBI and ENA identifiers for all genomes are provided in Additional file 2. Derived selection-audit tables, genomic annotations, marker calls, phylogenetic outputs, local-context evidence, and integron modules are supplied in Additional files 3 to 10. The underlying public sequence records remain available from NCBI and ENA under the listed accession numbers.

References

1. 1. Timm MR, Russell SK, Hultgren SJ. Urinary tract infections: pathogenesis, host susceptibility and emerging therapeutics. Nat Rev Microbiol. 2025;23:72–86. doi:10.1038/s41579-024-01092-4.
2. 2. Cao X, Cavaco LM, Lv Y, Li Y, Zheng B, Wang P, et al. Molecular characterization and antimicrobial susceptibility testing of *Escherichia coli* isolates from patients with urinary tract infections in 20 Chinese hospitals. J Clin Microbiol. 2011;49:2496–2501. doi:10.1128/JCM.02503-10.
3. 3. Van Goethem N, Descamps T, Devleesschauwer B, Roosens NHC, Boon NAM, Van Oyen H, et al. Status and potential of bacterial genomics for public health practice: a scoping review. Implement Sci. 2019;14:79. doi:10.1186/s13012-019-0930-2.
4. 4. Rempel OR, Laupland KB. Surveillance for antimicrobial resistant organisms: potential sources and magnitude of bias. Epidemiol Infect. 2009;137:1665–1673. doi:10.1017/S0950268809990100.
5. 5. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 2006;60:1136–1151. doi:10.1111/j.1365-2958.2006.05172.x.
6. 6. Stoesser N, Sheppard AE, Pankhurst L, De Maio N, Moore CE, Sebra R, et al. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. mBio. 2016;7:e02162-15. doi:10.1128/mBio.02162-15.
7. 7. Pitout JDD, Peirano G, Chen L, DeVinney R, Matsumura Y. *Escherichia coli* ST1193: following in the footsteps of *E. coli* ST131. Antimicrob Agents Chemother. 2022;66:e00511-22. doi:10.1128/aac.00511-22.
8. 8. Li F, Ye K, Li X, Ye L, Guo L, Wang L, et al. Genetic characterization of carbapenem-resistant *Escherichia coli* from China, 2015–2017. BMC Microbiol. 2021;21:248. doi:10.1186/s12866-021-02307-x.
9. 9. Xia C, Yan R, Liu C, Zhai J, Zheng J, Chen W, et al. Epidemiological and genomic characteristics of global *bla*NDM-carrying *Escherichia coli*. Ann Clin Microbiol Antimicrob. 2024;23:58. doi:10.1186/s12941-024-00719-x.
10. 10. Liu YY, Wang Y, Walsh TR, Yi LX, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular

biological study. Lancet Infect Dis. 2016;16:161–168. doi:10.1016/S1473-3099(15)00424-7.

11. 11. Li R, Xie M, Zhang J, Yang Z, Liu L, Liu X, et al. Genetic characterization of *mcr-1*-bearing plasmids to depict molecular mechanisms underlying dissemination of the colistin resistance determinant. J Antimicrob Chemother. 2017;72:393–401. doi:10.1093/jac/dkw411.
12. 12. Gillings MR. Integrons: past, present, and future. Microbiol Mol Biol Rev. 2014;78:257–277. doi:10.1128/MMBR.00056-13.
13. 13. Jiang Y, Zhang Y, Lu J, Wang Q, Cui Y, Wang Y, et al. Clinical relevance and plasmid dynamics of mcr-1-positive *Escherichia coli* in China: a multicentre case-control and molecular epidemiological study. Lancet Microbe. 2020;1:e24-e33. doi:10.1016/S2666-5247(20)30001-X.
14. 14. Acman M, Wang R, van Dorp L, Shaw LP, Wang Q, Luhmann N, et al. Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene *bla*NDM. Nat Commun. 2022;13:1131. doi:10.1038/s41467-022-28819-2.
15. 15. Shafiq M, Zeng M, Permana B, Bilal H, Huang J, Yao F, et al. Coexistence of *bla*NDM-5 and *tet*(X4) in international high-risk *Escherichia coli* clone ST648 of human origin in China. Front Microbiol. 2022;13:1031688. doi:10.3389/fmicb.2022.1031688.
16. 16. Zhao M, He J, Zhang R, Feng J, Deng Y, Zhang J. Epidemiological characteristics of New Delhi metallo-beta-lactamase-producing Enterobacteriaceae in the Fourth Hospital of Hebei Medical University. BMC Infect Dis. 2023;23:298. doi:10.1186/s12879-023-08242-8.
17. 17. Wang R, van Dorp L, Shaw LP, Bradley P, Wang Q, Wang X, et al. The global distribution and spread of the mobilized colistin resistance gene *mcr-1*. Nat Commun. 2018;9:1179. doi:10.1038/s41467-018-03205-z.
18. 18. Zhang X, Wan W, Yu H, Wang M, Zhang H, Lv J, et al. New Delhi metallo-beta-lactamase 5-producing *Klebsiella pneumoniae* sequence type 258, Southwest China, 2017. Emerg Infect Dis. 2019;25:1209–1213. doi:10.3201/eid2506.181939.
19. 19. Ba X, Guo Y, Moran RA, Doughty EL, Liu B, Yao L, et al. Global emergence of a hypervirulent carbapenem-resistant *Escherichia coli* ST410 clone. Nat Commun. 2024;15:494. doi:10.1038/s41467-023-43854-3.
20. 20. Wang C, Zhang H, Zhao R, Tsui CKM, Deng S. Genomic characterization and antimicrobial resistance of ESBL-producing, *Escherichia coli* isolates in Suzhou, China. Infect Drug Resist. 2025;18:1049–1057. doi:10.2147/IDR.S488794.
21. 21. Boutzoukas AE, Komarow L, Chen L, Hanson B, Kanj SS, Liu Z, et al. International epidemiology of carbapenemase-producing *Escherichia coli*. Clin Infect Dis. 2023;77:499–509. doi:10.1093/cid/ciad288.
22. 22. Gong L, Tang N, Chen D, Sun K, Lan R, Zhang W, et al. A nosocomial respiratory infection outbreak of carbapenem-resistant *Escherichia coli* ST131 with multiple transmissible *bla*KPC-2-carrying plasmids. Front Microbiol. 2020;11:2068. doi:10.3389/fmicb.2020.02068.
23. 23. Liu Y, Wang Q, Qi T, Zhang M, Chen R, Si Z, et al. Molecular epidemiology of *mcr-1*-positive polymyxin B-resistant *Escherichia coli* producing extended-spectrum beta-lactamase (ESBL) in a tertiary hospital in Shandong, China. Pol J Microbiol. 2024;73:363–375. doi:10.33073/pjm-2024-032.

24. 24. Haines MEK, Hodges FE, Nale JY, Mahony J, van Sinderen D, Kaczorowska J, et al. Analysis of selection methods to develop novel phage therapy cocktails against antimicrobial resistant clinical isolates of bacteria. Front Microbiol. 2021;12:613529. doi:10.3389/fmicb.2021.613529.
25. 25. Duan Y, Gao H, Zheng L, Liu S, Cao Y, Zhu S, et al. Antibiotic resistance and virulence of extraintestinal pathogenic *Escherichia coli* (ExPEC) vary according to molecular types. Front Microbiol. 2020;11:598305. doi:10.3389/fmicb.2020.598305.
26. 26. Fang Y, Tao S, Chen H, Xu Y, Chen L, Liang W. Influence of bacterial morphotype on urine culture and molecular epidemiological differences in *Escherichia coli* harboring bacterial morphotype-induced urinary tract infections. Microbiol Spectr. 2025;13:e00980-24. doi:10.1128/spectrum.00980-24.
27. 27. McNally A, Oren Y, Kelly D, Pascoe B, Dunn S, Sreecharan T, et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. PLoS Genet. 2016;12:e1006280. doi:10.1371/journal.pgen.1006280.
28. 28. Wang Y, He J, Sun L, Jiang Y, Hu L, Leptihn S, et al. IS26 mediated *bla*CTX-M-65 amplification in *Escherichia coli* increase the antibiotic resistance to cephalosporin in vivo. J Glob Antimicrob Resist. 2023;35:202–209. doi:10.1016/j.jgar.2023.09.018.
29. 29. He Q, Huang J, Li B. Draft genome sequences of three *Escherichia coli* sequence type 131 strains. Microbiol Resour Announc. 2019;8:e01474-18. doi:10.1128/MRA.01474-18.
30. 30. Xu T, Gong Y, Su X, Zhu P, Dai J, Xu J, et al. Phenome-genome profiling of single bacterial cell by Raman-activated gravity-driven encapsulation and sequencing. Small. 2020;16:2001172. doi:10.1002/smll.202001172.
31. 31. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884-i890. doi:10.1093/bioinformatics/bty560.
32. 32. Shen W, Le S, Li Y, Hu F. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 2016;11:e0163962. doi:10.1371/journal.pone.0163962.
33. 33. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–1075. doi:10.1093/bioinformatics/btt086.
34. 34. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257. doi:10.1186/s13059-019-1891-0.
35. 35. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. Microb Genom. 2018;4:e000192. doi:10.1099/mgen.0.000192.
36. 36. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, 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. doi:10.1038/s41598-021-91456-0.
37. 37. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–2069. doi:10.1093/bioinformatics/btu153.
38. 38. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2024;42:1303–1312. doi:10.1038/s41587-023-01953-y.

39. 39. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 2017;33:3340–3347. doi:10.1093/bioinformatics/btx433.
40. 40. Neron B, Littner E, Haudiquet M, Perrin A, Cury J, Rocha EPC. IntegronFinder 2.0: identification and analysis of integrons across bacteria, with a focus on antibiotic resistance in *Klebsiella*. Microorganisms. 2022;10:700. doi:10.3390/microorganisms10040700.
41. 41. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018;4:e000206. doi:10.1099/mgen.0.000206.
42. 42. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020;21:180. doi:10.1186/s13059-020-02090-4.
43. 43. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–1534. doi:10.1093/molbev/msaa015.
44. 44. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole-genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15. doi:10.1093/nar/gku1196.
45. 45. Xu J, He F. Characterization of a NDM-7 carbapenemase-producing *Escherichia coli* ST410 clinical strain isolated from a urinary tract infection in China. Infect Drug Resist. 2019;12:1555–1564. doi:10.2147/IDR.S206211.
46. 46. Xu J, Guo H, Li L, He F. Molecular epidemiology and genomic insights into the transmission of carbapenem-resistant NDM-producing *Escherichia coli*. Comput Struct Biotechnol J. 2023;21:847–855. doi:10.1016/j.csbj.2023.01.004.
47. 47. Qin J, Wang Z, Xu H, Li Y, Zhou J, Yaxier N, et al. IncX3 plasmid-mediated spread of *bla*NDM gene in Enterobacteriaceae among children in China. J Glob Antimicrob Resist. 2024;37:199–207. doi:10.1016/j.jgar.2024.03.021.
48. 48. Zhao F, Feng Y, Lu X, McNally A, Zong Z. Remarkable diversity of *Escherichia coli* carrying *mcr-1* from hospital sewage with the identification of two new *mcr-1* variants. Front Microbiol. 2017;8:2094. doi:10.3389/fmicb.2017.02094.
49. 49. Jiang H, Ran M, Wang X, Chen Q, Wang J, Ruan Z, et al. Prevalence and characterization of class I integrons in multidrug-resistant *Escherichia coli* isolates from humans and food-producing animals in Zhejiang Province, China. BMC Microbiol. 2025;25:76. doi:10.1186/s12866-025-03794-y.

Figures

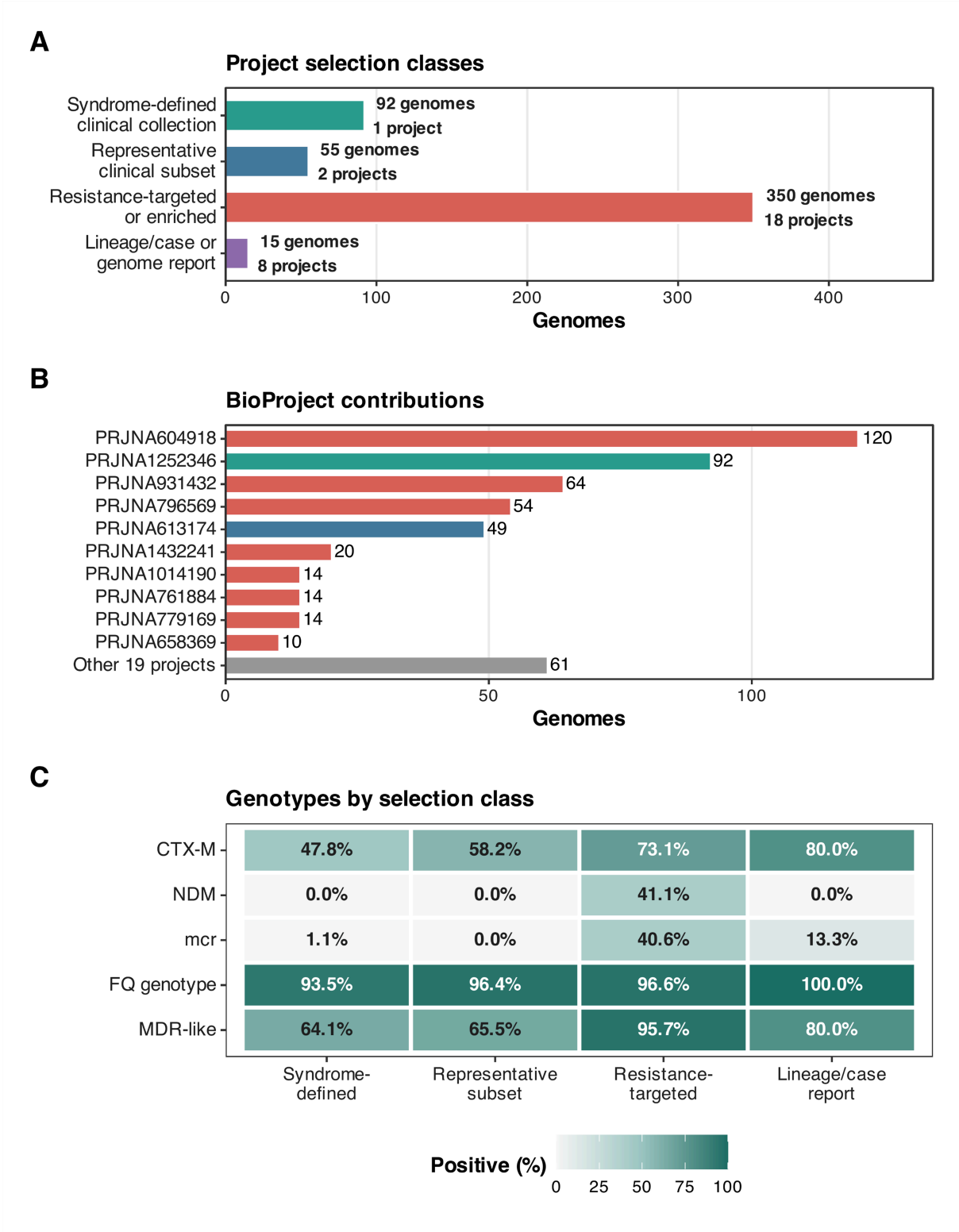


Figure 1

Selection mechanisms and genotype composition of contributing BioProjects

A, Number of genomes and BioProjects in each project selection class. **B**, Contributions of the ten largest BioProjects; the remaining 19 projects are grouped in grey. Colors for individual projects show selection

class. **C**, Resistance-genotype proportions within each selection class. Class assignments were based on original publications when available and otherwise on public BioProject or case records.

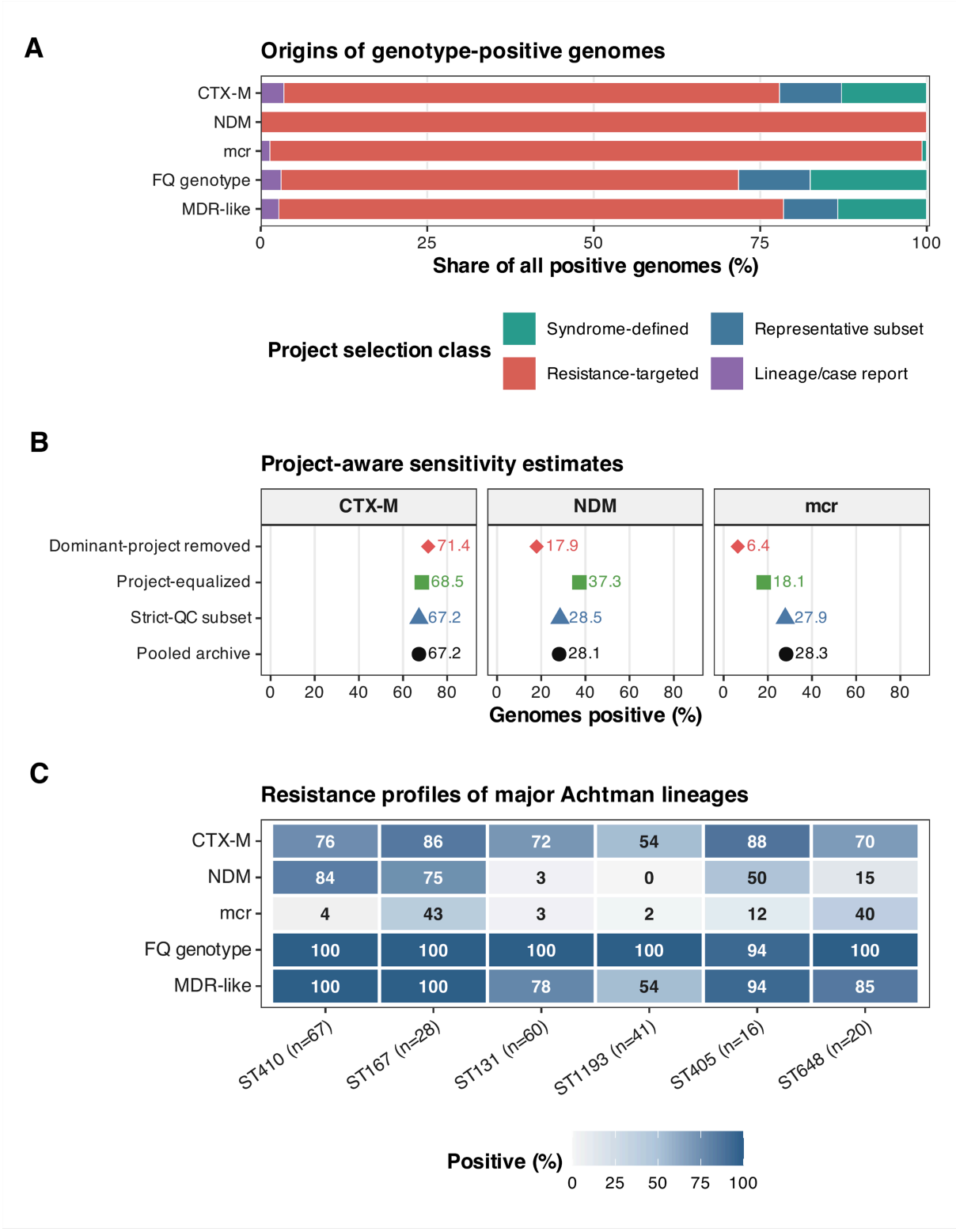


Figure 2

Selection-aware resistance signals and major Achtman lineage profiles

A, Selection-class origins of genomes positive for each resistance-genotype category. Each bar sums to the total number of positive genomes for that marker. **B**, Pooled archive, strict-quality, project-equalized, and dominant-project-exclusion estimates for CTX-M, NDM, and mcr. Project-equalized values are unweighted means of BioProject-specific proportions. **C**, Resistance-genotype proportions in six major Achtman lineages. FQ genotype denotes a fluoroquinolone resistance-associated genotype. MDR-like denotes markers in at least three antimicrobial categories.

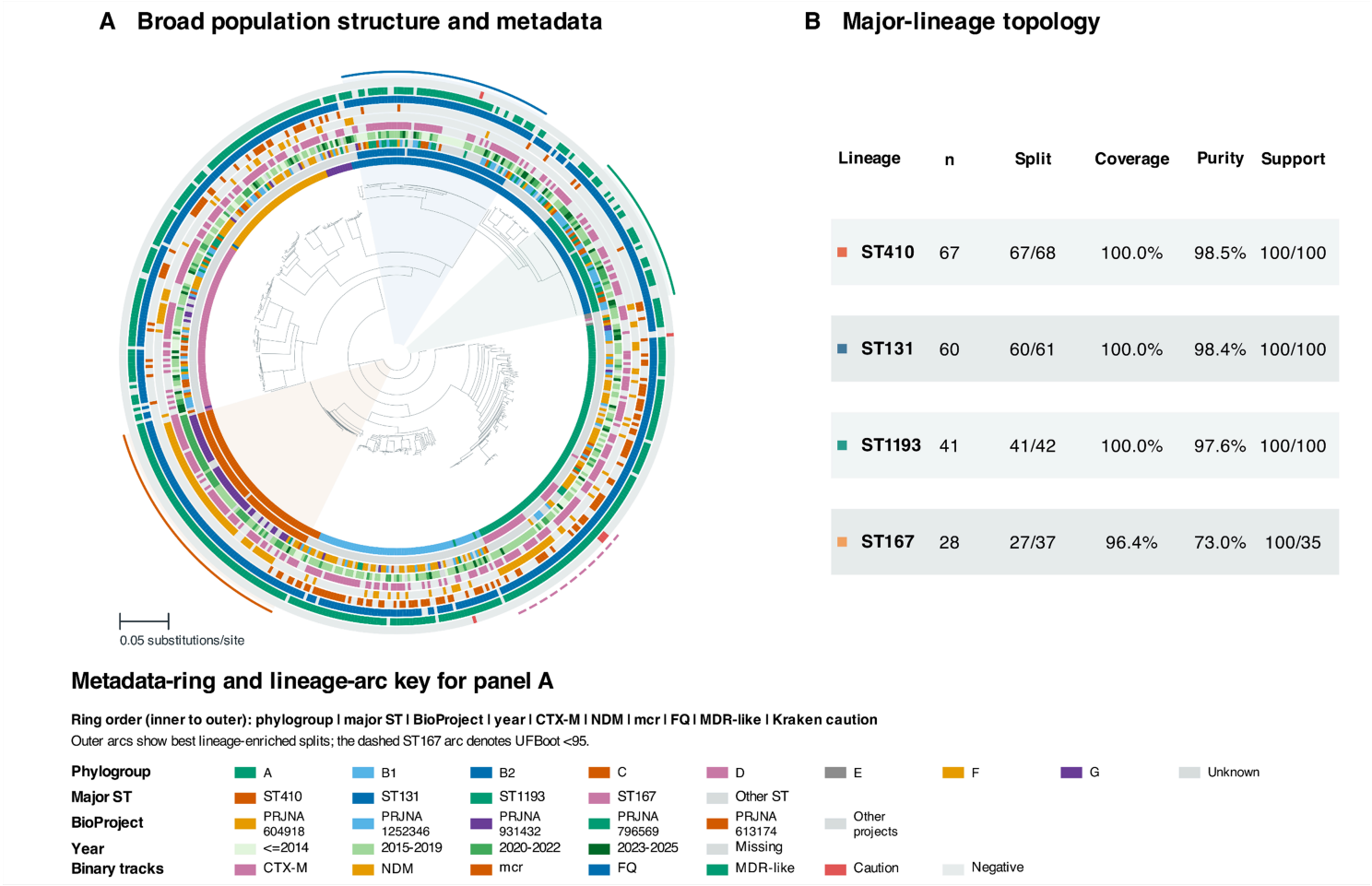


Figure 3

Broad population structure of the 512-genome archive cohort

A, Unrooted maximum-likelihood tree inferred from 329,173 Panaroo-derived variable sites under GTR+F+ASC+R10 with 1,000 ultrafast-bootstrap and 1,000 SH-aLRT replicates. Concentric tracks, from inner to outer, show phylogroup, major Achtman ST, dominant BioProject group, collection-year group, CTX-M, NDM, mcr, fluoroquinolone resistance-associated genotype, MDR-like genotype, and Kraken2 caution status. Colored outer arcs mark the best splits enriched for ST410, ST131, ST1193, and ST167; the dashed ST167 arc denotes UFBoot support below 95. The key immediately beneath panel A gives the metadata-ring order, colors, and lineage-arc notation. **B**, Coverage, purity, and node support for the best split representing each of the four major lineages. Panel A provides lineage-level context for genotype distribution.

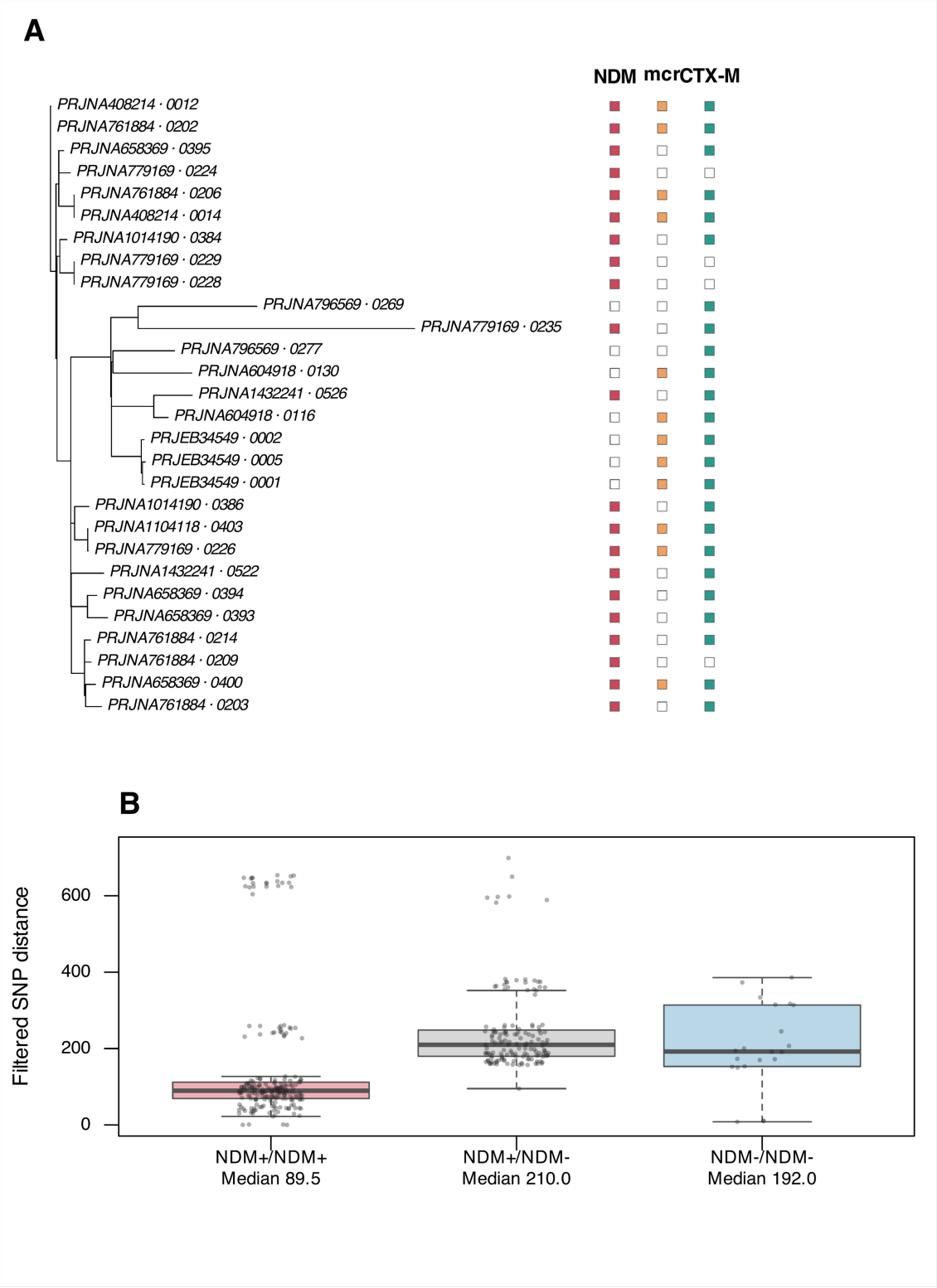


Figure 4

ST167 recombination-filtered phylogeny and pairwise SNP distances

A, Unrooted 28-tip ST167 maximum-likelihood tree inferred from 1,742 Gubbins-filtered polymorphic sites under TVMe+ASC. Filled squares mark NDM-, mcr-, and CTX-M-positive genomes. Tip labels combine BioProject and the numeric part of the cohort identifier. **B**, Pairwise filtered SNP distances grouped by NDM

status of the two genomes. Points show genome pairs, boxes show the interquartile range, and group labels give medians. Pairwise observations share genomes and were summarized descriptively.

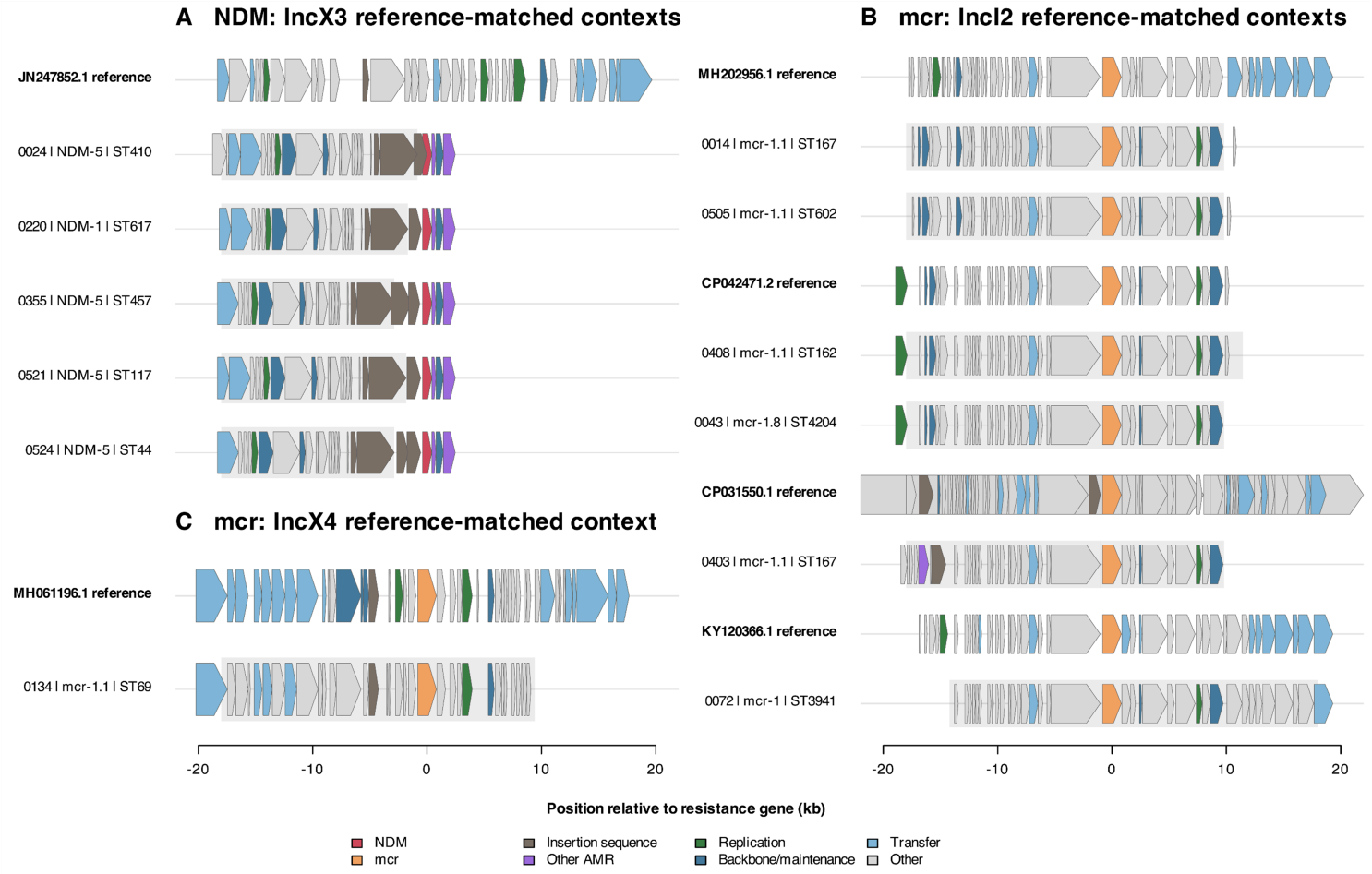


Figure 5

Representative NDM- and mcr-associated local contexts

A, Five nonredundant NDM local contexts aligned to the IncX3 plasmid JN247852.1. **B**, Six nonredundant mcr local contexts representing IncI2 reference matches to MH202956.1, CP042471.2, CP031550.1, and KY120366.1. **C**, One mcr-1.1 context from ST69 aligned to the IncX4 plasmid MH061196.1. All panels use the same horizontal scale, with the resistance gene centered at zero. Arrows show annotated coding sequences, and shaded blocks show regions selected for local comparison. Labels identify reference sequences or representative cohort genomes, resistance-gene variants, and Achtman sequence types. The panels show local contig structure.

Supplementary Files

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

- AdditionalFile7SoftwareDatabasesandParameters.xlsx

- AdditionalFile6IntegronCassetteModules.xlsx
- AdditionalFile4PopulationStructureandST167.xlsx
- AdditionalFile2CohortAccessionLedger.xlsx
- AdditionalFile3BioProjectSelectionAuditandSensitivity.xlsx
- AdditionalFile5NDMandmcrLocalContexts.xlsx
- AdditionalFile10ST167PhylogenyOutputs.zip
- AdditionalFile9BroadPhylogenyOutputs.zip
- AdditionalFile1SupplementaryInformation.docx
- AdditionalFile8ExploratoryFisherTests.xlsx