genes

MDPI

Article

Deposited-Assembly Availability, Raw-Read Rescue, and Cohort Composition in a Global *Klebsiella pneumoniae* Collection

Jianhua Li [1,2,3,†], Jing Guo [1,2,3,†], Xinxin Wang [1,2,3] and Jing Chen [1,2,3,*]

[1] Department of Critical Care Medicine, The First Affiliated Hospital of Xiamen University, Xiamen 361003, China
[2] School of Medicine, Xiamen University, Xiamen 361102, China
[3] Xiamen First Clinical Medical College, Fujian Medical University, Xiamen 361003, China
* Correspondence: dongqingfei@126.com
† These authors contributed equally to this work.

Abstract

Background/Objectives: Genome-input availability may affect which records enter studies of public bacterial genomes. We examined whether raw-read rescue changed the composition of a global human-clinical *Klebsiella pneumoniae* collection. **Methods**: We screened 150,089 National Center for Biotechnology Information (NCBI) records, of which 44,278 were eligible. We compared 10 prespecified metadata and genomic variables using adjusted association tests, effect sizes, incremental proportional changes, geographic and temporal stratification, and BioProject-level sensitivity analyses. **Results**: The final cohort comprised 42,589 genomes: 40,093 deposited assemblies and 2496 raw-derived replacements. Of 1687 eligible records absent from the historical input manifest, 1670 were used in a separate study and excluded from this analysis; 17 had no analyzable input recorded in this study. Raw-derived genomes accounted for 5.86% of the cohort and differed most clearly in lineage and geography. Changes in overall proportions were modest despite marked differences in several lineages. All records mapped uniquely to 340 BioProjects; two projects contributed 89.5% of raw-derived genomes. Sequence-type (ST) and K-locus differences persisted after joint removal of the two largest contributors of reconstructed genomes and in mixed-project comparisons with minimum source-group sample sizes. O-locus differences were smaller and more sensitive to sparse projects. **Conclusions**: Genome-input pathways were associated with cohort composition. Public-genome studies should report input provenance, study allocation, fallback yield, common post-assembly criteria, and submitting-project contributions. This study evaluates cohort composition rather than the accuracy of reconstructed genotypes and does not establish causal technical effects of reconstruction.

Keywords: *Klebsiella pneumoniae*; microbial genomics; public genome archives; assembly quality; raw-read reconstruction; selection bias; genomic surveillance

Check for updates

Academic Editors: Silvia Turroni and Ruiguang Ge

Received: 10 October 2026
Revised: 14 November 2026
Accepted: 16 November 2026
Published: 18 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Whole-genome sequencing supports bacterial population genetics and pathogen surveillance [1,2]. In *Klebsiella pneumoniae*, genomic analyses resolve sequence types, capsule and O-antigen loci, and resistance and virulence determinants [3,4]. Public genomes extend these analyses across settings, but their composition reflects surveillance priorities, sequencing capacity, and submission practices [5,6]. These sources of variation need to be considered when public genomes are reused for comparative studies. We selected

K. pneumoniae as a case study because published surveillance studies document genomic diversity across settings and established typing frameworks distinguish major lineage and antigen-locus groups [2–4].

An assembly accession does not establish suitability for a common analytical workflow. Public bacterial assemblies vary in quality; fragmentation and other defects can limit their use [7,8]. Eligible records with unavailable or unsuitable deposited assemblies may be recovered by reconstructing genomes from linked raw reads. Here, assembly usability refers to retention as an analyzable input in the study manifest. An assembly-only workflow can therefore exclude records that remain recoverable through another input pathway.

Genome recovery and dereplication address different aspects of cohort construction. Dereplication reduces repeated representation of duplicate or closely related genomes [9], whereas raw-read fallback can retain eligible records without suitable deposited assemblies. Public *K. pneumoniae* collections show uneven representation of resistance phenotypes and geographic sources [10,11]. Recovery yield alone does not indicate whether fallback changes the lineages and metadata profiles represented in an analytical cohort.

We compared retained deposited assemblies with replacements reconstructed from raw reads in a global human-clinical *K. pneumoniae* collection. We quantified fallback yield, compared the input groups across 10 prespecified variables, and measured the incremental change in cohort composition after successful reconstruction. Geographic, temporal, and BioProject analyses assessed the consistency of these differences. The aim was to evaluate selection during cohort construction by considering recovery yield together with the composition of recovered records, rather than to establish causal effects of reconstruction.

2. Materials and Methods

2.1. Study Design and Sampling Frame

We retrospectively examined deposited-assembly usability and raw-read fallback in a global collection of public *K. pneumoniae* records. Retained inputs were classified as deposited public assemblies or raw-read-derived replacements. We audited eligible records absent from the historical input manifest, distinguishing records used in another study from other records without an analyzable input in the study manifest. The primary comparison included genomes retained after species adjudication. Genome-input source was an observed characteristic, not an intervention or randomized exposure.

We reconstructed the sampling frame from the May 2026 species-specific National Center for Biotechnology Information (NCBI) RefSeq and GenBank assembly-summary files [12,13]. RefSeq and GenBank assembly summaries provided a consistent accession-level starting frame for linking assemblies and metadata; ENA was used separately as the source of fallback reads. Records with the organism name *K. pneumoniae* were considered initially. We resolved paired RefSeq and GenBank entries and multiple assemblies linked to the same BioSample. We then retained traceable current inputs and applied metadata, human clinical, and geographic eligibility criteria. Environmental, food-associated, animal-associated, laboratory, non-target, obsolete, duplicate, and other non-human records were excluded. Screening flags could overlap and were not treated as mutually exclusive reasons for exclusion. A separately curated abscess-enriched collection was developed for a phenotype-directed project with different case ascertainment criteria and a different genome-input recovery strategy. Records used in that project were excluded from the present analysis, and the separate collection was not pooled with the final analytical cohort.

2.2. Deposited-Assembly Assessment and Raw-Read Fallback

The initial screen identified eligible records with a deposited assembly listed in the public repository. Genome assemblies and raw reads were retrieved using scripts, download

utilities, and manual downloads from repository webpages. We first attempted to download and assess deposited assemblies. When the download was unsuccessful, or the assembly did not meet the input criteria, we attempted reconstruction from associated raw reads. Fallback could remain unsuccessful when raw reads could not be obtained or when no replacement assembly was retained. Download procedures varied across records.

We assigned the source groups from the historical input manifest: deposited assembly or raw-read-derived replacement. Species adjudication was applied before comparison. We reconciled archived processing attempts against the manifest to distinguish unsuccessful raw-read attempts that still yielded cohort membership via a deposited assembly from records absent from the manifest. To audit public-assembly retrieval, we distinguished terminal download failures, file-level exclusions after retrieval, and later quality measures. We report counts only for the historical batch for which these stages could be reconstructed.

Raw reads used for fallback reconstruction were retrieved from the European Nucleotide Archive (ENA) [14]. The historical reconstruction workflow used fastp 1.3.3 [15] for read preprocessing and Shovill 0.9.0 (https://github.com/tseemann/shovill, accessed on 9 October 2026) with SPAdes 3.15.5 [16] for assembly. Of the 2497 records in the historical reconstructed set, 2486 had recoverable fastp commands specifying paired inputs (-i and -I), paired trimmed outputs (-o and -O), paired-end adapter detection (–detect_adapter_for_pe), four threads (-w 4), and HTML and JSON reports (–html and –json). Shovill commands were recovered for all 2497 records and specified paired trimmed reads (–R1 and –R2), an output directory (–outdir), –cpus 4, and –force. For the 11 records without complete fastp commands, the recovered SPAdes calls used –only-assembler, –threads 4, and –memory 8; the k-mer lists varied by run. The archived processing script did not specify a minimum read-quality or sequencing-depth threshold for proceeding to assembly. We traced available input and output paths and used ENA annotations to characterize the associated runs (Table S6).

Recovered commands established correspondence between input and output paths; they were not used to infer successful completion of every processing step. The two sources were compared using the same downstream genotype categories. Data handling used Python v3.10.20. Sequence types were assigned with Kleborate v3.1.3 [4] and MLST v2.33.1 using PubMLST schemes [17]. K and O loci were assigned with Kaptive v3.2.1 [18]. Software-version evidence and code for the post-assembly analyses are included in the archived materials.

2.3. Assembly Quality Assessment

Archived assembly reports included contig counts, total assembly length and N50. The historical records did not establish uniform enforcement of the contig-count, N50, genome-length, GC-content or ambiguous-base thresholds across both sources. The recovered public-assembly input screen checked file origin and integrity indicators, including minimum file size. That screen was distinct from biological assembly quality assessment.

For quality control (QC), we measured contig count and N50 using contigs ≥500 bp in the archived assemblies linked to the locked cohort. We also assessed genome completeness and contamination with CheckM2 v1.1.0 [19], DIAMOND v2.1.11, and CheckM2 database v3. We linked estimates by accession, source and input file hash. We reused 170 estimates with matching full-file SHA-256 values and generated estimates for the remaining 42,419 genomes. Matching the files used for this audit does not establish that every file was unchanged from the historical workflow.

For the QC sensitivity analyses, we recalculated the 10 global Cramér's V estimates using the original categories and source labels. We evaluated contig count ≤500 and N50 ≥20 kb, and separately evaluated contamination ≤5% with completeness ≥90% or ≥95%.

The 95% completeness cutoff tested a stricter threshold. We then applied the contiguity and CheckM2 criteria jointly. We also recalculated the ST4843 odds ratio and repeated joint removal of the two largest raw-derived BioProjects in the quality-filtered sets (Table S9). The original 42,589-genome cohort remained the primary analysis. The contig-count and N50 cutoffs were used in the QC sensitivity analyses. We retained the historical cohort for the primary comparison because our objective was to describe the composition of that cohort. These cutoffs do not establish the accuracy of sequence-type or antigen-locus assignments.

2.4. Comparison Variables

We compared 10 prespecified variables between assembly-source groups: continent, country group, collection-period group, specimen type, infection-site group, clinical or colonization status, sampling module, sequence-type group, K-locus group, and O locus. The sequence-type hierarchy distinguished exact numeric assignments, locus-variant profiles, and unresolved results. Missing values were not imputed; they were retained in an explicit UNKNOWN_OR_MISSING category to keep category totals traceable. The variables describe three dimensions of cohort composition: geographic and temporal coverage, clinical and sampling context, and lineage and antigen-locus representation. A recovered metadata-standardization script classified records from keyword rules and retained an unknown/other fallback when a specific category could not be assigned. The exact version of the final metadata builder and any independent validation of these assignments could not be verified.

2.5. Statistical Analysis

For each variable, the two source groups were compared in a 2 × k contingency table using Pearson's chi-square test without Yates' correction. Association strength was summarized by Cramér's V. Values below 0.05 were classified as negligible, 0.05 to <0.10 as small, 0.10 to <0.20 as moderate, and 0.20 or greater as large. These categories were descriptive guides, not biological cutoffs. Analyses with any expected cell count below 5 were flagged. Each category was also compared with all remaining categories using a two-sided Fisher exact test and an odds ratio. Benjamini–Hochberg (BH) correction [20] was applied separately to the 10 global tests and to the pooled family of 110 category-specific Fisher tests across the 10 variables. Several high-dimensional tables were sparse. The tests did not account for dependence among genomes within a BioProject, and multiple-testing adjustment does not address that dependence. Interpretation prioritized Cramér's V and consistency across stratified and project-level checks; adjusted p values provided supporting evidence. Analyses used SciPy v1.17.1.

We retained the original category definitions. For the six global tables with an expected count below 5, a post hoc fixed-margin Monte Carlo test recalculated the Pearson statistic in 9999 simulated tables (seed 4651494; add-one p values). These checks address the asymptotic approximation for sparse tables; they do not account for within-BioProject dependence. We also recalculated the ten global Cramér's V values after excluding a flagged reconstruction and, separately, all 11 records with incomplete fastp command provenance (Table S7). These exclusions were sensitivity analyses and did not redefine the primary cohort.

2.6. Incremental and Stratified Analyses

We assessed changes in the combined collection after reconstructed genomes were added. For each variable, category proportions were calculated before and after their addition, and the largest absolute change was recorded in percentage points (pp). Total-variation distance expressed in pp (TVDpp) was calculated using the equation below. Here,

pcombined,i and passembly,i denote category proportions in the combined and assembly-only collections, respectively. TVDpp ranges from 0 for identical distributions to 100 for non-overlapping distributions. Sequence-type (ST), K-locus, and O-locus comparisons were repeated within geographic and collection-period strata containing at least 100 public-assembly genomes and 30 reconstructed genomes. p values from the 27 estimable stratified tests were adjusted together. Interpretation focused on effect size and consistency. These tests did not account for within-BioProject clustering, so adjusted p values were considered supportive evidence.

$$TVD_{pp} = 100 \times 0.5 \times \sum_{i} \left| p_{combined,i} - p_{assembly,i} \right|$$

Exploratory comparisons assessed the metadata composition of eligible records represented in versus absent from the historical input manifest (Table S2). Manifest membership was treated as a separate endpoint from the outcome of an individual raw-read processing attempt. These comparisons do not isolate technical assembly failure. In a separate descriptive audit (Table S8), we linked the 1670 records allocated to the other study and the 42,591 retained historical records to the locked 44,278-record metadata snapshot by exact versioned assembly accession. We compared recorded country, collection period, and broad specimen categories derived by the same exploratory text rules in both groups. Percentages used all records in each group; ambiguous or unspecified sources were kept separate. To quantify the shift associated with study allocation, we compared the pooled metadata distribution of these two groups (n = 44,261) with that of the retained group. The 17 other absent records were tabulated separately.

2.7. BioProject Sensitivity Analyses

BioProject identifiers were recovered by exact assembly-accession linkage to the eligible-record metadata. In these sensitivity analyses, each exact numeric ST was treated as a category, with locus-variant profiles and unresolved assignments pooled separately. Exact K- and O-locus labels were taken from the final master table, including explicit categories for missing labels. These definitions differ from the grouped ST and K-locus categories in the main analysis. Effect sizes were compared within the sensitivity analyses and did not replace the main grouped estimates. We tabulated project contributions and removed 17 large projects individually. These comprised the union of the ten largest projects overall and the ten largest contributors of raw-derived genomes. We also removed the two largest contributors of reconstructed genomes simultaneously.

Further analyses were restricted to BioProjects contributing both input sources. Exploratory post hoc checks required at least 10 or 50 records per source within a project. Category-frequency distributions were calculated for each source in each eligible project and averaged with equal project weights; the same weights were used for both sources. Cramér's V and source-to-source total-variation distance (TVD, 0–1) summarized these distributions. This TVD compares the two input groups and differs from the incremental combined-versus-public TVDpp in Section 2.6. Exploratory percentile intervals were calculated from 1000 bootstrap resamples of whole projects with seed 20261010. They were not interpreted as null-hypothesis tests. The small number of projects, sparse categories, and change in the weighting target limit inference. Source assignments and sample counts were held fixed, and no new genomic endpoints were introduced (Table S5). In a separate post hoc analysis of the seven projects with at least 10 records of each source, we used the main-analysis groupings for ST and K locus and the original O-locus categories. We summarized source differences with the median and range of the seven project-specific TVDs. For each endpoint, we summed the project-specific Pearson statistics and compared the sum with 9999 independent fixed-margin draws within projects (seed 4651494; add-one

p values, BH across three endpoints; Table S7). This analysis conditions on BioProject and assumes exchangeability within projects under the null. It cannot account for unmeasured within-project clustering or sampling differences.

BioProject analyses used Python and accession-linked source tables. The Data Availability Statement identifies the source data, code, and numerical outputs.

3. Results

3.1. Cohort Formation

The source files contained 150,089 assembly records: 30,642 RefSeq and 119,447 GenBank entries. After paired-record resolution and metadata, traceability, human-clinical, and geographic screening, 44,278 eligible records remained (Figure 1; Table 1). The historical input manifest contained 40,094 deposited assemblies and 2497 raw-derived assemblies, totaling 42,591 analyzable inputs (96.19% of eligible records). Of the 1687 eligible records absent from the manifest, 1670 were used in the separate phenotype-directed study and excluded from this analysis. The other 17 had no retained analyzable input recorded for this study. In the traceable early download batch (42,606 accessions), three lacked a successful public assembly download at the end of FTP rescue. Of the 42,603 recorded downloads, 11 were excluded by the 1,000,000-byte input gate and one because it was a test file. Two of the 17 had been skipped at the early overlap screen. Record-level raw-read failure could not be established for all 17; the historical batch counts cannot be generalized to the entire retrieval workflow.

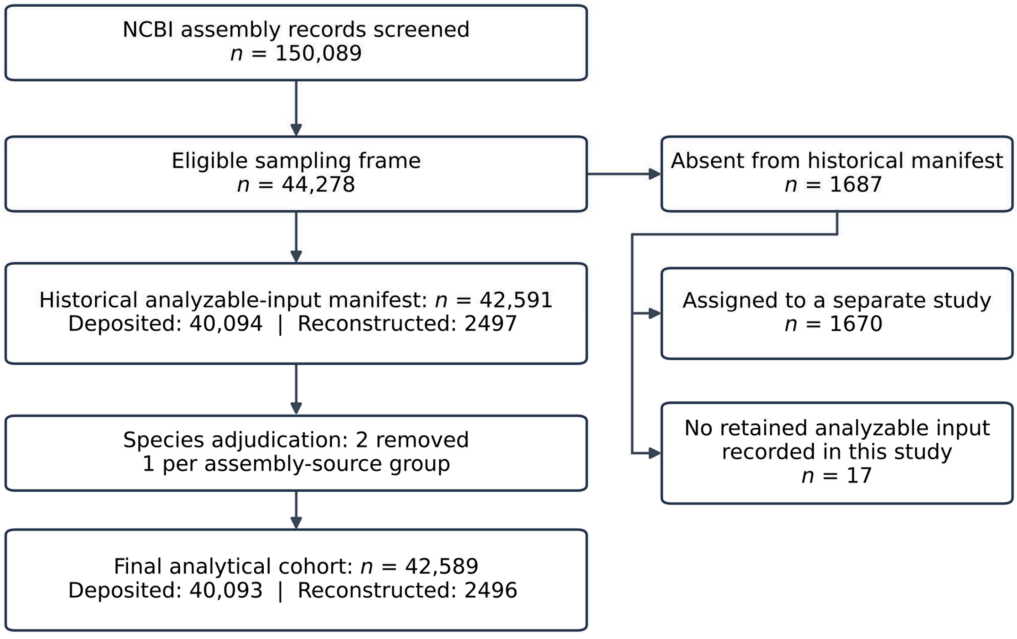


Figure 1. Cohort formation and disposition of records absent from the historical manifest. Study allocation (1670) is shown separately from records without a retained analyzable input (17). The latter count describes manifest membership; it is not a count of verified raw-read download failures. NCBI, National Center for Biotechnology Information.

Compared with the 42,591 historical retained records, the 1670 allocated records more often had an abscess/pus specimen label (720/1670 [43.11%] vs. 118/42,591 [0.28%]) or were recorded from China (696/1670 [41.68%] vs. 11,048/42,591 [25.94%]). A smaller share was recorded from the United States (22.28% vs. 35.64%) or collected in 2023 or later (22.34% vs. 25.81%). The abscess/pus share was 1.89% in the pooled metadata set of 44,261 records and 0.28% among historical retained records, a decrease of 1.62 percentage

points. The pooled set includes records for which genome usability in this study was not assessed (Table S8).

Table 1. Cohort formation and contribution of raw-read fallback.

Stage	*n*	% of Eligible Frame	Role
Eligible sampling frame	44,278	100.00	Upstream geographically resolved human-clinical records
Public assembly retained before species adjudication	40,094	90.55	Historical deposited-assembly inputs
Raw-derived assembly retained before species adjudication	2497	5.64	Historical reconstructed inputs
Historical analyzable-input manifest	42,591	96.19	40,094 public + 2497 raw-derived
Eligible records outside historical manifest	1687	3.81	1670 allocated to another study; 17 without retained input
Used in another phenotype-directed study	1670	3.77	Not included in the present analytical cohort
Other records without retained input	17	0.04	Historical retrieval or input-screen outcomes; see Table S2
Excluded by species adjudication	2	<0.01	One public and one raw-derived input
Final public-assembly group	40,093	90.55	94.14% of the final analytical cohort
Final raw-derived group	2496	5.64	5.86% of the final analytical cohort
Final analytical cohort	42,589	96.19	100.00% of the final analytical cohort

The percentage column uses the eligible sampling frame (n = 44,278) throughout. Final source shares, shown in the last column for the final groups, use the final analytical cohort (n = 42,589). Nested disposition counts and successive processing stages are not additive exclusions.

Species adjudication excluded two genomes, one from each input source. The final cohort contained 42,589 genomes: 40,093 public assemblies and 2496 raw-derived replacements. Raw-derived replacements accounted for 5.86% of the cohort.

All 2497 run accessions associated with the historical reconstructed set were annotated by ENA as Illumina paired-end. Recovered command logs linked fastp and Shovill paths for 2486 records. Among the 11 records without complete fastp commands, nine logs showed gzip decompression errors, and two showed unexpected end-of-file errors. All 11 were retained in the reconstructed group and had Shovill command records, SPAdes completion markers and QUAST v5.3.0 reports [8]. Shovill logs also recorded error exits, and the available records did not confirm successful read preprocessing (Table S6). All 11 archived assemblies met the completeness (≥90%) and contamination (≤5%) criteria. Ten also met the contiguity criteria (≤500 contigs and N50 ≥ 20 kb); the remaining assembly had 672 contigs and an N50 of 15,094 bp (Tables S6 and S9). Excluding the flagged record alone or all 11 records changed Cramér's V by less than 0.001 in each of the ten global comparisons (Table S7).

3.2. Composition by Genome-Input Source

Nine of the 10 global tests had BH-adjusted p values below 0.05 (Figure 2; Table 2). Sequence-type group had the strongest association with assembly source (Cramér's V = 0.382; adjusted $p < 0.001$), followed by country group (V = 0.209; adjusted $p < 0.001$). Associations were moderate for continent (V = 0.161), K-locus group (V = 0.134), and collection period (V = 0.125). Interpretation prioritized Cramér's V because of the large sample size and unmodeled BioProject clustering; adjusted p values provided supporting evidence.

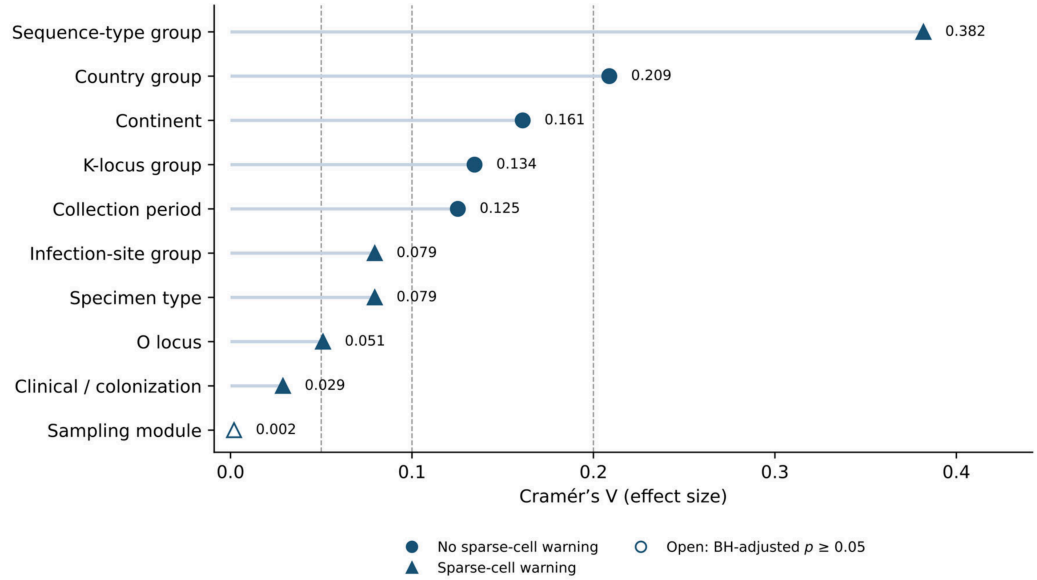


Figure 2. Effect sizes for compositional differences by assembly source. Horizontal position and numeric labels show Cramér's V for 40,093 deposited and 2496 reconstructed assemblies. Dashed lines denote study-defined descriptive cutoffs (0.05, 0.10 and 0.20). Triangles flag at least one expected count below 5; circles indicate no such flag. Filled symbols have BH-adjusted $p < 0.05$; open symbols have BH-adjusted $p \geq 0.05$. BH, Benjamini–Hochberg.

Table 2. Global compositional comparisons between retained public-assembly and raw-read-derived replacement groups.

Endpoint	k	Retained Public Assembly, *n*	Raw-Read-Derived Replacement, *n*	BH-Adjusted *p*	Cramér's V	Magnitude	Sparse-Cell Warning
Continent	6	40,093	2496	<0.001	0.161	Moderate	No
Country group	21	40,093	2496	<0.001	0.209	Large	No
Collection-year group	5	40,093	2496	<0.001	0.125	Moderate	No
Specimen type	9	40,093	2496	<0.001	0.079	Small	Yes
Infection-site group	9	40,093	2496	<0.001	0.079	Small	Yes
Clinical/colonization status	4	40,093	2496	<0.001	0.029	Negligible	Yes
Sampling module	3	40,093	2496	0.9232	0.002	Negligible	Yes
Sequence-type group	23	40,093	2496	<0.001	0.382	Large	Yes
K-locus group	21	40,093	2496	<0.001	0.134	Moderate	No
O locus	9	40,093	2496	<0.001	0.051	Small	Yes

BH, Benjamini–Hochberg; ST, sequence type. Cramér's V summarizes the global 2 × k association. All *p* values are two-sided and were adjusted across the 10 prespecified global tests. Magnitude labels are prespecified descriptive aids and do not replace endpoint-specific interpretation. A sparse-cell warning indicates at least one expected count below 5. These tests did not account for within-BioProject clustering; adjusted *p* values are supporting evidence, and effect-size estimates are prioritized. k denotes the number of categories; K and O denote the capsule and O-antigen loci, respectively.

3.3. Other Compositional Differences

Associations were small for specimen type and infection-site group (V = 0.079 for both) and O locus (V = 0.051). The association with clinical or colonization status was negligible (V = 0.029). Sampling-module composition did not differ (adjusted $p = 0.9232$; V = 0.002). Comparisons with sparse expected counts were flagged and interpreted using effect size

and consistency across strata. Fixed-margin simulation gave the same nominal 0.05 decisions as the original analysis for the six flagged global tables (Monte Carlo $p = 0.8835$ for sampling module and ≤ 0.0008 for the other five). These p values do not account for project clustering (Table S7).

3.4. Sequence-Type Classification

An exact numeric ST was assigned to 41,592 genomes, a locus-variant profile to 959, and an unresolved result to 38. Together, these categories accounted for all 42,589 genomes (Table S1).

Category-level comparisons showed marked lineage imbalance. ST4843 accounted for 270/2496 (10.82%) reconstructed genomes and 61/40,093 (0.15%) deposited genomes (odds ratio, 79.60; 95% confidence interval (CI), 60.09–105.44). Locus-variant profiles were enriched in the reconstructed group (13.42% vs. 1.56%; odds ratio, 9.81), as was ST45 (6.21% vs. 2.85%; odds ratio, 2.26). ST11, ST258, and ST147 were substantially underrepresented among reconstructed genomes (2.64% vs. 18.52%, 0.28% vs. 10.35%, and 0.60% vs. 7.38%, respectively; Table S1).

3.5. Effect on the Combined Collection

Reconstructed genomes accounted for 5.86% of the final collection. Changes in the combined proportions were smaller than the differences between source groups. The largest change in a single category was an increase of 1.75 pp for North America. Total-variation distance was 2.59 pp for the ST group, 2.35 for the country group, and 1.94 for the continent (Table S3). The marked differences between input groups thus corresponded to modest changes in the overall collection.

3.6. Stratified Comparisons

The ST-group association remained significant after correction in all nine estimable geographic or temporal strata (Figure 3; Table S4). Cramér's V ranged from 0.189 to 0.503; eight strata had large effects, and one had a moderate effect. The largest values occurred in records collected in 2023 or later (V = 0.503) and in North America (V = 0.484). South America and the period before 2010 did not meet the minimum reconstructed-sample threshold. K-locus associations were moderate or large in all estimable strata. O-locus associations were mostly small or negligible and were not significant in Europe or Oceania. Associations with input source were stronger and more consistent for sequence type and K locus than for O locus.

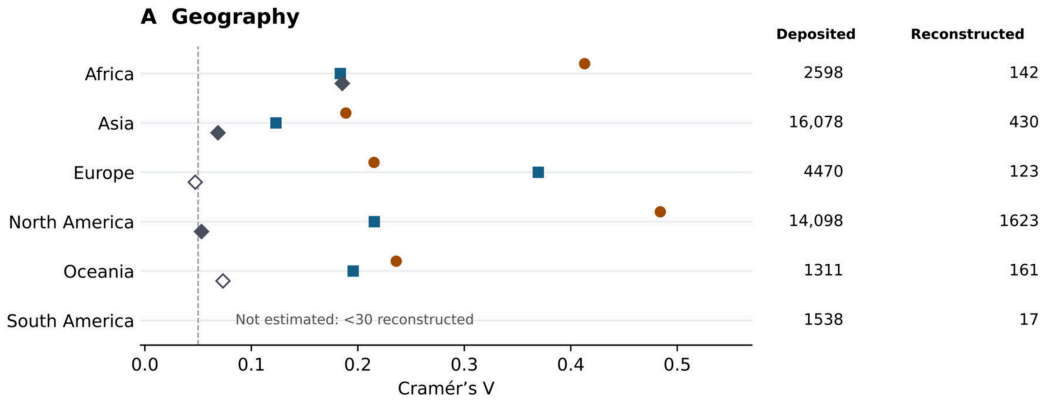


Figure 3. *Cont.*

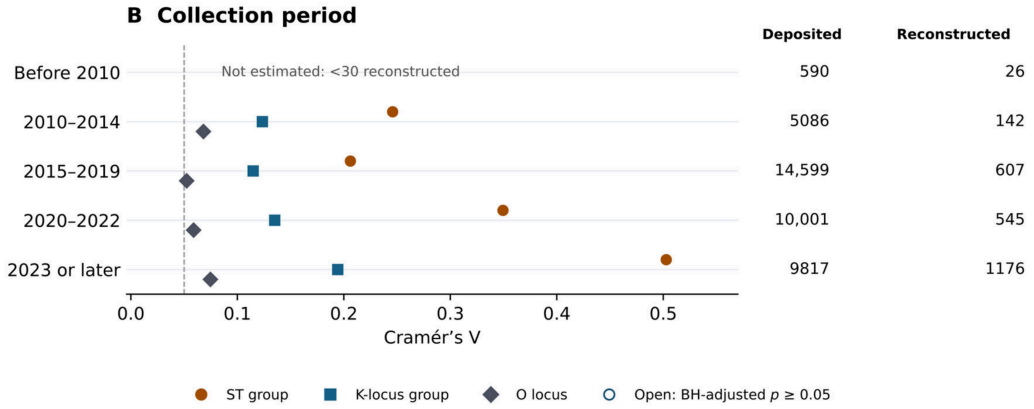


Figure 3. Stratified lineage and antigen-locus comparisons. The two columns show the numbers of deposited and reconstructed genomes in each geographic (**A**) or temporal (**B**) stratum. Shapes and colors distinguish sequence-type (ST) group, K-locus group and O locus. Filled symbols have BH-adjusted $p < 0.05$; open symbols have BH-adjusted $p \geq 0.05$. The dashed line at Cramér's V = 0.05 is a descriptive effect-size cutoff. South America (17 reconstructed genomes) and the period before 2010 (26) did not meet the minimum of 100 deposited and 30 reconstructed genomes. Expected counts below 5 are listed in Table S4. BH, Benjamini–Hochberg.

Reconstructed genomes were concentrated in particular regions and collection periods. Of 2496 genomes, 1623 (65.02%) originated from North America, and 1176 (47.12%) were collected in 2023 or later. These patterns may reflect overlapping project contributions and sequencing practices; they do not identify causal effects of reconstruction.

3.7. BioProject Sensitivity Results

All 42,589 genomes were linked uniquely to one eligible-record metadata entry and one BioProject, with no missing or conflicting assignments. The cohort represented 340 projects: 333 contributed public assemblies, 20 contributed raw-derived genomes, and 13 contributed both. PRJNA514245 contributed 1299 raw-derived genomes (52.0%), and PRJNA288601 contributed 935 (37.5%). Together, they accounted for 89.5% of the raw-derived group. With PRJNA224116 (160 genomes), the three largest contributors accounted for 95.9% (Table S5; Figure S1). Using the added exact-category definitions, full-cohort Cramér's V values were 0.579 for ST, 0.176 for K locus, and 0.051 for O locus. Across the 17 individual project removals, the corresponding ranges were 0.551–0.629, 0.164–0.212, and 0.050–0.055. After joint removal of PRJNA514245 and PRJNA288601, 338 projects contributed 22,642 deposited and 262 reconstructed genomes. The exact-category Cramér's V values were 0.648 for ST, 0.214 for K locus and 0.034 for O locus. PRJNA224116 still contributed 160/262 (61.1%) of the reconstructed genomes in this subset.

The seven mixed-source projects with at least 10 records per source retained 28,017 public assemblies and 2458 raw-derived genomes (98.5% of the raw-derived group). Equal-project-weighted source-to-source TVDs were 0.716 for ST, 0.322 for K locus, and 0.056 for O locus. Restriction to the three projects with at least 50 records per source retained 27,484 public assemblies and 2394 raw-derived genomes (95.9%). The corresponding TVDs were 0.662, 0.313, and 0.037. ST and K-locus differences persisted in these descriptive comparisons; O-locus differences were smaller and sensitive to projects with sparse source groups. Table S5 provides exploratory cluster-bootstrap intervals, not significance tests. The differences were not confined to one dominant project. These checks cannot, however, separate input-source associations from project sampling or within-project processing. Concentration of raw-derived genomes in a few projects limits generalization beyond this collection. In the within-project analysis using the main-analysis categories, median project-specific TVD was

0.491 for the ST group (range 0.324–0.643), 0.423 for the K-locus group (0.207–0.566), and 0.105 for the O locus (0.035–0.449). The three observed statistics exceeded all 9999 simulated sums (BH-adjusted Monte Carlo $p = 0.0001$ for each; Table S7).

3.8. Assembly Quality Sensitivity

All 42,589 genomes had a linked CheckM2 estimate. Median completeness was 100% in both groups (interquartile range, 100–100%). Median contamination was 0.52% (0.21–1.21%) for deposited assemblies and 0.21% (0.12–0.58%) for raw-derived assemblies. Of the currently archived assemblies, 215 did not satisfy at least one contiguity threshold: 169 of 40,093 deposited assemblies and 46 of 2496 raw-derived assemblies (Table S9A).

Applying only the contig-count and N50 filters retained 42,374 genomes (39,924 deposited and 2450 reconstructed). The maximum absolute change in Cramér's V across the 10 comparisons was 0.00194 (Table S9A,B). Joint filtering retained 41,714 genomes (39,278 deposited and 2436 raw-derived) at the 90% completeness threshold and 41,681 (39,245 deposited and 2436 raw-derived) at 95%. Across the 10 comparisons, the maximum absolute changes in Cramér's V were 0.0073 and 0.0087, respectively. Figure S2 shows the ST-group, K-locus-group and O-locus effect sizes before and after joint filtering. The ST4843 odds ratios were 85.42 (95% CI, 63.97–114.04) and 85.34 (95% CI, 63.92–113.95), compared with 79.60 in the original cohort (Table S9C). The two largest raw-derived BioProjects still accounted for 2197/2436 genomes (90.19%) in both subsets. In the joint 90% subset after removing both projects, exact-category V was 0.667 for ST, 0.224 for K locus and 0.037 for O locus (22,299 deposited and 239 reconstructed genomes; Table S9D).

4. Discussion

4.1. Genome Recovery and Cohort Composition

Public-genome cohorts reflect both the records deposited in repositories and the inputs retained for analysis [1,5,6]. In this collection, the latter stage was associated with lineage and geographic composition even though fallback contributed a small fraction of genomes. Raw-read fallback retained 2496 genomes, or 5.86% of the final collection, whose deposited assemblies were not retained in the historical input manifest. Their composition differed from that of retained deposited assemblies, and the sequence-type association persisted in every estimable geographic and temporal stratum. The largest collection-wide change in a category was nevertheless 1.75 pp, and the ST-group TVD was 2.59 pp. A compositionally different recovered subset therefore had a modest effect on overall proportions.

These lineage and geographic differences are consistent with evidence that public collections are not neutral samples of pathogen populations. Kochan et al. found that public *K. pneumoniae* bloodstream-isolate genomes were enriched for antimicrobial resistance relative to consecutive isolates from one hospital [10]. Ekwanzala documented uneven geographic sampling in African-linked *K. pneumoniae* species-complex genomes [11]. Those studies examined submission and sampling rather than raw-read rescue. Our comparison concerns an additional selection stage: retention of deposited assemblies and recovery of eligible records through linked reads. It does not identify which submission, sequencing, quality, or curation mechanism produced the observed differences.

BioProject sensitivity analyses indicate that the ST and K-locus differences were not confined to one dominant project: both persisted after individual removal of large projects and in project-restricted comparisons. Two projects nevertheless contributed 89.5% of raw-derived genomes, making their sampling designs and processing histories central to interpretation. Work on the Salmonella UoWUCC collection illustrates how inclusion of differently sampled records can broaden the diversity represented in a public genomic resource [21]. Our checks do not establish that raw-read fallback has the same effect across

projects or improves population representativeness. Equal-project weighting and minimum-size restrictions describe source-group composition rather than technical equivalence. O-locus differences were smaller and more sensitive to sparse projects. Different project objectives can favor particular lineages, sites or time periods; comparisons within BioProjects reduce the influence of between-project differences but cannot rule out different sampling or processing of the two sources within a project. The concentration remained after quality filtering (90.19% of the raw-derived subset), so the QC results do not remove this limitation.

4.2. Interpreting Lineage and Geographic Differences

Effect sizes were more informative than statistical significance alone in this collection of more than 42,000 genomes. Several variables had adjusted p values below 0.001 but small or negligible Cramér's V values, whereas sequence type and geography showed larger differences. Interpretation therefore considered effect size, adjusted p value, and sparse-cell status together. Geographic and temporal stratification showed persistence of the sequence-type pattern, and BioProject analyses examined variation between submission projects. The original chi-square tests did not account for clustering within studies, so their p values may be optimistic. Persistence of ST and K-locus differences after individual large-project removal and minimum-size restrictions does not overcome the limits imposed by project concentration. The very small clinical or colonization and sampling-module effect sizes should not be interpreted as clinically important differences in this cohort.

The biological characteristics associated with *K. pneumoniae* lineages differ across settings, so altered lineage representation can affect interpretation of a reused genomic cohort. David et al. linked the European carbapenem-resistant epidemic to the spread of particular clonal lineages [22], whereas Wyres et al. described different lineage distributions and resistance–virulence profiles in South and Southeast Asian bloodstream isolates [23]. Surveys across human, animal, and environmental sources and clinical resistance-focused collections also describe high-risk or carbapenemase-associated lineages [24–26]. These studies provide context for the lineage imbalance observed here. They do not explain the input-source differences or establish that raw-read fallback changes resistance or virulence; those outcomes were not evaluated in this study.

Replacement assemblies were not a random subset of the eligible sampling frame. Enrichment of ST4843 and locus-variant profiles, together with depletion of several established epidemic clones, indicates that a small reconstructed fraction can matter for lineage-specific analyses. Sampling, submission, assembly-quality, and source-selection practices may contribute to these contrasts. Recovery also differs from dereplication: reducing redundant genomes [9] does not restore eligible records whose deposited inputs were not retained. Both decisions can affect representation at different stages of the workflow.

The lineage contrasts should not be interpreted as clinical phenotype associations. Genotyping frameworks treat lineage, antigen loci, and resistance or virulence features as distinct attributes [4], and resistance can depend on individual carbapenemase variants [27]. A lineage label alone cannot establish a resistance or virulence phenotype. Public metadata incompletely capture clinical syndromes, surveillance objectives, and project-level sampling decisions. ST4843 enrichment is a difference between input groups within this cohort, not evidence of an association with a clinical syndrome.

4.3. Implications for Public-Genome Studies

Reporting fallback yield together with the composition of recovered records would make selection during cohort construction easier to assess. Replacement assemblies accounted for approximately 6% of this collection and changed most overall proportions little, but altered the representation of several lineages and geographic sources. Stable

collection-wide proportions can mask shifts within a sequence type, antigen locus, or stratum. Recommendations for open pathogen genomic analysis emphasize traceable data and workflows [1]. In assembly-first studies, reporting should include the number of deposited assemblies not retained, reconstruction yield, the lineages and strata recovered, and the common post-assembly checks applied to both sources. This study examined selection during the database-to-analysis workflow, an endpoint that wet-laboratory validation would not directly test.

Selection during cohort construction needs to be distinguished from genotyping accuracy. Merda et al. showed that assembly tools and intrinsic genome composition can influence bacterial cgMLST profiles [28], and Lüth et al. examined sequencing-platform compatibility in Listeria monocytogenes genome comparisons [29]. Those studies support evaluation of technical comparability but do not quantify ST or K/O concordance in this collection. Kaptive 3 has been evaluated for in silico antigen typing [18]; use of a validated typing tool does not establish equivalence between deposited and reconstructed inputs. Uniform assembly checks, when verifiably applied, can improve comparability but do not distinguish sampling differences from potential assembly-dependent differences in typing. Paired assemblies from the same samples are needed to address this question. Future workflows should retain accession-level provenance, read-processing commands, assembler versions and parameters, and submitting-project identifiers.

Raw-read fallback cannot make a convenience sample representative of its source population. Uneven surveillance coverage and data sharing remain relevant when additional eligible records are recovered [1,6]. Without a population-based reference, we cannot determine whether replacement assemblies improved correspondence with disease incidence or prevalence. Within this collection, deposited-assembly usability was associated with composition, and successful fallback retained records enriched for particular lineages and geographic sources.

4.4. Limitations

Deposited-assembly retention and successful fallback were not randomized, so associations between source groups do not establish causal effects of reconstruction. The reconstructed group included only records whose deposited assemblies were not retained and whose replacements were generated successfully. It therefore does not represent all records lacking suitable deposited input. Study allocation also affected cohort membership: 1670 eligible records absent from the historical manifest were used in another study and excluded from this analysis. The phenotype-directed allocation changed the composition of the retained metadata set, particularly the representation of records labeled abscess/pus. This descriptive comparison does not estimate how including the allocated records would have changed genome usability or genotype composition. For the other 17 records, the historical evidence supports separate public-assembly retrieval or input-screen outcomes but does not establish record-level raw-read failure. Harmonization of public metadata also carried a risk of misclassification.

High-dimensional comparisons contained sparse cells, and rare categories can influence exact-ST effect sizes. The original global and geographic/temporal tests did not account for BioProject clustering, so nominal and adjusted p values may be optimistic. BioProject identifiers were recovered for all records, and project-level checks showed persistent ST and K-locus differences. Raw-derived genomes were nevertheless concentrated in a few projects. Only 13 projects contributed both sources; minimum-size restrictions reduced this number to seven or three. Equal-project weighting changes the target population and does not remove within-project differences in geography, time, sampling, sequencing technology, or assembly quality. Bootstrap intervals based on few projects remain exploratory and

do not establish a causal processing effect. The within-project simulation reduces dependence on between-project differences; its exchangeability assumption does not exclude correlation from unmeasured subgroups within a project. Intervals based on only three or seven resampled projects have limited resolution and should not be read as evidence of generalizability.

Some archived assemblies did not meet the contiguity thresholds used in the QC sensitivity analyses. The historical records do not establish whether these thresholds were applied uniformly at cohort entry. Source-group differences were similar after QC filtering. This finding addresses sensitivity to the evaluated filters; without paired assemblies from the same isolates, it cannot establish the accuracy or concordance of ST and K/O assignments. No independent validation cohort or population-based reference was available. The findings concern the composition of this collection rather than population prevalence or clinical outcomes.

5. Conclusions

Raw-read fallback retained eligible *K. pneumoniae* genomes whose composition differed from that of retained deposited public assemblies, particularly in lineage and geography. ST and K-locus differences were observed in geographic, temporal and BioProject comparisons, although raw-derived inputs were concentrated in a few projects. Modest changes in collection-wide proportions could mask larger shifts within lineages or strata. Studies reusing public genomes should report assembly retention, study allocation, fallback yield together with the composition of recovered records, accession-level provenance, and project contributions. Common post-assembly criteria should be applied to both sources. These findings describe associations within this collection and do not establish causal technical effects of reconstruction.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes17111456/s1, Table S1: Grouped sequence-type results; Table S2: Exploratory manifest-membership results and record disposition; Table S3: Incremental compositional impact of raw-derived genomes; Table S4: Geographic and temporal stratified results; Table S5: BioProject composition, individual large-project removal, mixed-project minimum-size checks, and equal-project-weighted effects with exploratory bootstrap intervals; Table S6: Sequencing-platform annotations and historical preprocessing-command provenance; Table S7: Sparse-table simulations, category-level Fisher tests, within-BioProject comparisons, and assembly-provenance sensitivity checks; Table S8: Metadata composition of records allocated to another study and historical retained records; Table S9: Assembly QC and sensitivity analyses; Figure S1: Project concentration and minimum-size sensitivity of source-to-source composition differences. Complete numeric source tables and the accession-level BioProject manifest accompany the Supplementary Materials. The complete 2497-record platform mapping and records with incomplete command provenance are supplied as machine-readable files with Table S6; Figure S2: Grouped effect sizes before and after joint quality filtering.

Author Contributions: Conceptualization, J.L., J.G. and J.C.; methodology, J.L., J.G., X.W. and J.C.; software, validation, formal analysis, data curation and visualization, J.L., J.G. and X.W.; investigation, J.L., J.G. and X.W.; writing—original draft preparation, J.L. and J.G.; writing—review and editing, X.W. and J.C.; supervision and project administration, J.C.; funding acquisition, J.L. and J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Xiamen Municipality (grant number 3502Z202373092) and the Medical and Health Guidance Project of Xiamen Municipality (grant number 3502Z20254ZD1079).

Institutional Review Board Statement: Institutional review board approval was not required because the study used only publicly available genomic records and associated metadata, with no intervention or direct contact with human participants.

Informed Consent Statement: Not applicable. This study used only publicly available genomic records and associated metadata.

Data Availability Statement: The accession-level manifests, source tables, BioProject mappings, figure data and analysis code for the locked cohort are associated with Zenodo version 1.6.0 (https://doi.org/10.5281/zenodo.23283049). Public assemblies and reads are available through their accession identifiers in the source repositories. The sensitivity analyses and accession-level quality estimates accompany Tables S7–S9. Historical processing records were incomplete; full fastp commands could not be recovered for 11 records in the historical fallback manifest.

Acknowledgments: The authors thank the investigators and institutions that deposited the public genome assemblies, sequence reads, and metadata used in this study. The authors remain responsible for the content and scientific conclusions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Black, A.; MacCannell, D.R.; Sibley, T.R.; Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat. Med.* **2020**, *26*, 832–841. [CrossRef]
2. Argimón, S.; David, S.; Underwood, A.; Abrudan, M.; Wheeler, N.E.; Kekre, M.; Abudahab, K.; Yeats, C.A.; Goater, R.; Taylor, B.; et al. Rapid genomic characterization and global surveillance of *Klebsiella* using Pathogenwatch. *Clin. Infect. Dis.* **2021**, *73*, S325–S335. [CrossRef]
3. Holt, K.E.; Wertheim, H.; Zadoks, R.N.; Baker, S.; Whitehouse, C.A.; Dance, D.; Jenney, A.; Connor, T.R.; Hsu, L.Y.; Severin, J.; et al. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, an urgent threat to public health. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3574–E3581. [CrossRef]
4. Lam, M.M.C.; Wick, R.R.; Watts, S.C.; Cerdeira, L.T.; Wyres, K.L.; Holt, K.E. A genomic surveillance framework and genotyping tool for *Klebsiella pneumoniae* and its related species complex. *Nat. Commun.* **2021**, *12*, 4188. [CrossRef]
5. Albright, S.; Louca, S. Trait biases in microbial reference genomes. *Sci. Data* **2023**, *10*, 84. [CrossRef]
6. Gardy, J.L.; Loman, N.J. Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* **2018**, *19*, 9–20. [CrossRef]
7. Wanchai, V.; Patumcharoenpol, P.; Nookaew, I.; Ussery, D. dBBQs: dataBase of Bacterial Quality scores. *BMC Bioinform.* **2017**, *18*, 483. [CrossRef]
8. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
9. Salamzade, R.; Kottapalli, A.; Kalan, L.R. skDER and CiDDER: Two scalable approaches for microbial genome dereplication. *Microb. Genom.* **2025**, *11*, 001438. [CrossRef]
10. Kochan, T.J.; Nozick, S.H.; Medernach, R.L.; Cheung, B.H.; Gatesy, S.W.M.; Lebrun-Corbin, M.; Mitra, S.D.; Khalatyan, N.; Krapp, F.; Qi, C.; et al. Genomic surveillance for multidrug-resistant or hypervirulent *Klebsiella pneumoniae* among United States bloodstream isolates. *BMC Infect. Dis.* **2022**, *22*, 603. [CrossRef]
11. Ekwanzala, M.D. Public genomic surveillance of African *Klebsiella pneumoniae* species complex reveals uneven sampling and high ESBL gene carriage. *Sci. Rep.* **2026**, *16*, 31383. [CrossRef]
12. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
13. Sayers, E.W.; Cavanaugh, M.; Frisse, L.; Pruitt, K.D.; Schneider, V.A.; Underwood, B.A.; Yankie, L.; Karsch-Mizrachi, I. GenBank 2025 update. *Nucleic Acids Res.* **2025**, *53*, D56–D61. [CrossRef]
14. O'Cathail, C.; Ahamed, A.; Burgin, J.; Cummins, C.; Devaraj, R.; Gueye, K.; Gupta, D.; Gupta, V.; Haseeb, M.; Ihsan, M.; et al. The European Nucleotide Archive in 2024. *Nucleic Acids Res.* **2025**, *53*, D49–D55. [CrossRef]
15. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
16. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]

17. Jolley, K.A.; Bray, J.E.; Maiden, M.C.J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **2018**, *3*, 124. [CrossRef]
18. Stanton, T.D.; Hetland, M.A.K.; Löhr, I.H.; Holt, K.E.; Wyres, K.L. Fast and accurate in silico antigen typing with Kaptive 3. *Microb. Genom.* **2025**, *11*, 001428. [CrossRef]
19. Chklovski, A.; Parks, D.H.; Woodcroft, B.J.; Tyson, G.W. CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **2023**, *20*, 1203–1212. Correction in *Nat. Methods* **2024**, *21*, 735. https://doi.org/10.1038/s41592-024-02248-z. [CrossRef]
20. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
21. Achtman, M.; Zhou, Z.; Alikhan, N.F.; Tyne, W.; Parkhill, J.; Cormican, M.; Chiou, C.S.; Torpdahl, M.; Litrup, E.; Prendergast, D.M.; et al. Genomic diversity of *Salmonella enterica*—The UoWUCC 10K genomes project. *Wellcome Open Res.* **2021**, *5*, 223. [CrossRef]
22. David, S.; Reuter, S.; Harris, S.R.; Glasner, C.; Feltwell, T.; Argimon, S.; Abudahab, K.; Goater, R.; Giani, T.; Errico, G.; et al. Epidemic of carbapenem-resistant *Klebsiella pneumoniae* in Europe is driven by nosocomial spread. *Nat. Microbiol.* **2019**, *4*, 1919–1929. [CrossRef]
23. Wyres, K.L.; Nguyen, T.N.T.; Lam, M.M.C.; Judd, L.M.; van Vinh Chau, N.; Dance, D.A.B.; Ip, M.; Karkey, A.; Ling, C.L.; Miliya, T.; et al. Genomic surveillance for hypervirulence and multi-drug resistance in invasive *Klebsiella pneumoniae* from South and Southeast Asia. *Genome Med.* **2020**, *12*, 11. [CrossRef]
24. Morgado, S.; Fonseca, E.; Vicente, A.C. Genomics of *Klebsiella pneumoniae* species complex reveals the circulation of high-risk multidrug-resistant pandemic clones in human, animal, and environmental sources. *Microorganisms* **2022**, *10*, 2281. [CrossRef]
25. Del Rio, A.; Fox, V.; Muresu, N.; Sechi, I.; Cossu, A.; Palmieri, A.; Scutari, R.; Alteri, C.; Sotgiu, G.; Castiglia, P.; et al. A whole-genome sequencing-based approach for the characterization of *Klebsiella pneumoniae* co-producing KPC and OXA-48-like carbapenemases circulating in Sardinia, Italy. *Microorganisms* **2023**, *11*, 2354. [CrossRef]
26. Rizvi, M.; Khan, N.; Fatima, A.; Bushra, R.; Zehra, A.; Saeed, F.; Gul, K. The microbiological characteristics and genomic surveillance of carbapenem-resistant *Klebsiella pneumoniae* isolated from clinical samples. *Microorganisms* **2025**, *13*, 1577. [CrossRef]
27. Fontana, C.; Favaro, M.; Campogiani, L.; Malagnino, V.; Minelli, S.; Bossa, M.C.; Altieri, A.; Andreoni, M.; Sarmati, L. Ceftazidime/avibactam-resistant *Klebsiella pneumoniae* subsp. *pneumoniae* isolates in a tertiary Italian hospital: Identification of a new mutation of the carbapenemase type 3 (KPC-3) gene conferring ceftazidime/avibactam resistance. *Microorganisms* **2021**, *9*, 2356. [CrossRef]
28. Merda, D.; Vila-Nova, M.; Bonis, M.; Boutigny, A.L.; Brauge, T.; Cavaiuolo, M.; Cunty, A.; Regnier, A.; Sayeb, M.; Vingadassalon, N.; et al. Unraveling the impact of genome assembly on bacterial typing: A One Health perspective. *BMC Genom.* **2024**, *25*, 1059. [CrossRef]
29. Lüth, S.; Fuchs, J.; Deneke, C. Compatibility of whole-genome sequencing data from Illumina and Ion Torrent technologies in genome comparison analysis of *Listeria monocytogenes*. *Microb. Genom.* **2025**, *11*, 001389. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.