BacTaxID: A universal framework for standardized bacterial classification

Miguel D. Fernández-de-Bobadilla,[1,2], Val F. Lanza*[3,4].

1. Institute of Functional Biology and Genomics, CSIC & University of Salamanca, Salamanca, Spain.
2. Microbiology Department, Ramón y Cajal University Hospital and Ramón y Cajal Health Research Institute (IRYCIS), Madrid, Spain;
3. Translational Genomics (NGS) and Bioinformatics Unit; Ramón y Cajal Health Research Institute (IRYCIS), Madrid, Spain.
4. Biomedical Research Center Network of Infectious Diseases (CIBERINFEC); Madrid, Spain.

Abstract

Bacterial strain typing is key to surveillance, outbreak investigation and microbial ecology, yet current systems remain species-specific, reference-dependent and lack a universal, interpretable metric of genomic relatedness. Here, we introduce BacTaxID, a fully configurable, whole-genome k-mer–based framework that encodes each genome as a numeric sketch and organizes strains into hierarchical clusters with user-defined similarity thresholds. BacTaxID distances are strictly proportional to Average Nucleotide Identity (ANI), providing a direct quantitative link between vectorial typing and genome-wide divergence. Applied to 2.3 million genomes from All the Bacteria across 67 genera, BacTaxID demonstrates universal concordance species and sub-species classification systems, while capturing finer strain-level diversity than traditional reference-based approaches. In simulated surveillance and real outbreak datasets, BacTaxID reproduces SNP- and cgMLST-based definitions while enabling rapid, scalable screening. Precomputed genus-level schemes and an open implementation provide a practical, genus-agnostic alternative to classical typing systems for standardized bacterial classification.

Introduction

Bacterial typing establishes a standardized framework for strain characterization that enables effective communication among researchers, clinicians, and public health officials. Over a century, bacterial typing has evolved from serology to whole-genome sequencing (WGS) methods that now offer scalable, reproducible strain classification. Typing is indispensable for epidemiology and surveillance, enabling outbreak detection, transmission tracking, and recognition of emerging clones with increased virulence or resistance.[1–3]

Modern practice predominantly relies on multilocus sequence typing (MLST) and its genome-scale derivatives, core-genome and whole-genome MLST (cgMLST/wgMLST). MLST targets 6–9 housekeeping genes to define portable Sequence Types (STs)[4], cgMLST in a larger dataset of core-genes and wgMLST captures accessory variation for maximal resolution[5]. Distributed cgMLST[6] (dcgMLST) schemes use hash functions to assign alleles unique identifiers, eliminating centralized nomenclature databases but maintaining database dependency for comparative analysis.

However, fundamental limitations persist. Typing schemes are species-specific with no universal system applicable across bacteria[7–9]. MLST and cgMLST identifiers are arbitrary numerical designations that lack biological interpretation, ST10 and ST11 may be phylogenetically distant, while closely related strains may have different STs[8,10]. Moreover, cgMLST distances saturate more rapidly than SNP distances, limiting utility to within-species comparisons, because allelic profiles across a fixed core-gene set quickly converge to maximal divergence once genomes are moderately distant, whereas genome-wide SNP counts continue to scale with deeper evolutionary separation and thus retain discriminatory power at broader taxonomic ranges[11,12].

Hierarchical systems including HierCC[13] and LIN[14–16] organize strains across multiple nested levels but generate unwieldy multi-level nomenclatures. Labels such as HC2850_2221_1895_1565_1301_982_759_501 break the communicative simplicity of classical MLST identifiers (e.g. ST131), making them awkward for reporting and representing a step back from short, portable sequence-type labels Both systems rely on single-linkage clustering susceptible to chaining artifacts, where intermediate genomes bridge phylogenetically distant lineages, obscuring epidemiologically meaningful structure.

The substantial diversity in mutation rates, recombination frequencies, and population structures among bacterial species precludes universal thresholds for delineating transmission chains[17–19]. Contemporary frameworks remain constrained to species-specific implementations requiring resource-intensive empirical validation, fragmenting the typing landscape into isolated, non-interoperable systems[3,7,20]. What is critically needed is a hierarchical universal typing system that computes strain relatedness using a single metric applicable across all bacteria, then organizes isolates into hierarchical similarity levels from which users select appropriate resolution thresholds[8,13,14,16]. Crucially, this system should provide compact strain labels that transparently encode hierarchical membership, enabling immediate interpretation without database lookups.

Here we introduce BacTaxID, a universal, genome-based typing system designed to overcome these limitations. BacTaxID employs whole-genome k-mer profiling and

implements a novel pseudo-clique-based clustering algorithm that organizes isolates hierarchically. Unlike single-linkage methods vulnerable to chaining[21], our approach enforces that each genome satisfies the minimum similarity threshold to a defined fraction of cluster members, preventing phylogenetically distant lineages from merging and providing robustness to outliers. The architecture maintains consistent performance across millions of genomes, essential for global-scale surveillance. K-mer-based distance metrics eliminate scheme-dependent biases through reference-free profiling, where each genome's unique signature emerges from its k-mer content without external reference comparisons, ensuring portable, comparable outputs and eliminating the need for curated locus databases. BacTaxID generates compact, interpretable identifiers combining the communicative simplicity of traditional MLST with hierarchical resolution, enabling efficient communication while preserving full structural information. This flexibility extends BacTaxID's utility from routine surveillance to high-resolution outbreak delineation. To demonstrate capabilities, we analyzed 2.3 million genomes spanning 67 genera and 3,926 species from the All the Bacteria database[22]. Complete results, pre-computed classification schemes for all 67 genera, and interactive tools for classification exploration are freely available at www.bactaxid.org, providing the community with comprehensive access to generated datasets and enabling seamless integration of BacTaxID classifications into existing bioinformatics workflows.

Results

Hierarchical Typing Algorithm

BacTaxID implements a graph-theoretic hierarchical clustering approach for automated bacterial genome classification, integrating sketch-based distance estimation with iterative level-by-level cluster assignment for scalable nomenclature generation.

BacTaxID employs Binwise Densified MinHash[23,24] combined with ntHash[25,26] for efficient k-mer hashing. Each genome undergoes reference-free k-mer decomposition to generate compact sketch representations, which are then condensed into a unique hash-based identifier specific to each sequence's genomic content. Pairwise distances are estimated through Jaccard similarity approximation via merge-sort comparison of ordered hash values. This sketch-based approach achieves computational efficiency orders of magnitude greater than alignment-dependent methods while maintaining correlation with evolutionary distances. The hierarchical search strategy restricts the comparison space: at level L_0, the algorithm compares the query genome againt all classifiers; at subsequent levels (L_i, $i > 0$), searches are confined to classifiers sharing identical typing codes at preceding levels. This decomposition reduces computational complexity from $O(N)$ in flat schemes to $O(\log N)$ or $O(\sqrt{N})$ in hierarchical frameworks, enabling tractable processing as databases expand.

BacTaxID's sketch-based representation shares a fundamental conceptual similarity with cgMLST and related allele-calling systems: both encode genomic information as fixed-length numeric vectors where each position represents a distinct genomic feature. In cgMLST, sketch_sizevector elements correspond to specific genes, each encoded as a unique allele number; in BacTaxID, vector elements represent k-mer hash bins of size *sketch_size*. However, BacTaxID achieves equivalent vectorial encoding with significant methodological advantages: the reference-free k-mer approach eliminates dependence on predefined locus databases, and pairwise vector comparisons directly correlate with genome-wide evolutionary distance measured by Average Nucleotide Identity (ANI), providing a quantitative link between typing results and phylogenetic relationships.

At each hierarchical level, reference sketches are retrieved and distances to the query genome are computed. If the minimum distance falls below the level-specific threshold, the query inherits the reference typing code. The algorithm then evaluates whether the query exhibits threshold-compliant similarity with a minimum proportion of reference sequences (*click_threshold* parameter). If this criterion is satisfied and cluster size remains below the maximum capacity (*reference_size* parameter), the query becomes a "classifier" genome. Queries meeting similarity requirements but joining size-saturated clusters receive "satellite" status. Only classifiers serve as references for subsequent assignments, while satellite genomes contribute exclusively to forming novel clusters at higher hierarchical levels. This stratification prevents outliers, hybrid strains, or hypermutants from perturbing the classification framework while allowing atypical strains to establish independent clusters when sufficiently abundant.

When assignment criteria are not met, the system constructs a distance graph and applies *clique* detection algorithms to identify maximal complete subgraphs, sets of genomes where

all pairwise distances satisfy the threshold. Valid *cliques* satisfying minimum size and density requirements establish novel cluster groups with hierarchical codes propagated to all members. The algorithm enforces hierarchical consistency through monotonicity constraints, ensuring codes at level L_{i+1} refine groupings from level L_i.

BacTaxID produces a unified DuckDB[27] database file consolidating all typing information: scheme configuration parameters, hierarchical strain identifiers, network clustering metadata, and the complete k-mer sketch database. This self-contained format ensures portability and reproducibility while enabling seamless interoperability with third-party tools through DuckDB's API ecosystem (SQL, Python, R, command-line access). The system architecture supports both static classification of new isolates according to pre-established schemes and dynamic scheme updates through iterative re-clustering as genomic databases expand.

All algorithmic parameters, including the number of hierarchical levels, level-specific similarity thresholds, k-mer length, *click_threshold*, and *reference_size*, are user-configurable during scheme initialization, enabling adaptation to diverse bacterial species and epidemiological contexts. This approach combines multi-level resolution with alignment-free distance estimation, eliminating dependency on predefined loci catalogs characteristic of cgMLST systems, and distinguishes it from traditional agglomerative clustering approaches susceptible to chaining artifacts.

BacTaxID implements a semi-centralized typing architecture that separates reference-free genomic profiling from standardized nomenclature assignment. Genome signature generation, the creation of each sequence's unique k-mer hash representation, is inherently decentralized and reference-independent, enabling local, autonomous computation without centralized infrastructure. Conversely, assignment of hierarchical typing codes requires centralized scheme maintenance to preserve nomenclature consistency across global surveillance networks[28]. This architectural design, supported by open-source implementation and portable database formats, enables truly distributed profiling while maintaining comparable, portable classifications, substantially reducing adoption barriers while ensuring universal standardization.

All The Bacteria dataset.

To demonstrate BacTaxID's capabilities, we analyzed the All the Bacteria (ATB) database[22], a large-scale, open-access project that assembles and catalogs millions of bacterial and archaeal genomes from public repositories. The ATB collection comprised 2.4 million quality-controlled genomes as of August 2024, providing a standardized resource for comparative genomic analysis. We selected all genera containing more than 500 genomes, yielding a comprehensive dataset of 2.3 million genomes distributed across 67 genera and 3,926 species.

Individual typing schemes were generated for each genus using identical parameterization: six hierarchical clustering levels based on average nucleotide identity (ANI) thresholds of 96%, 98%, 99%, 99.5%, 99.9%, and 99.99%. Consequently, each genome is assigned a hierarchical typing code composed of up to six integers separated by periods (e.g.,

1.3.1.8.12.1), where each integer represents the cluster assignment at a successive level of resolution. All schemes are publicly available at https://doi.org/10.5281/zenodo.17226627, and comprehensive information including the complete dataset, metadata from All the Bacteria, classification results, and interactive exploration tools are freely available at www.bactaxid.org. This unified platform provides researchers with seamless access to all BacTaxID classification data, enabling integration with external analyses and supporting both surveillance applications and evolutionary comparative studies. The uniform framework establishes a universal typing architecture that supports species-specific strain classification for surveillance, epidemiological investigation, and outbreak detection, while simultaneously enabling comparative analysis of population structure across phylogenetically diverse bacterial lineages, thereby extending BacTaxID's utility from clinical microbiology to microbial ecology and evolutionary biology.

We selected the genus as the operational taxonomic level for scheme construction due to ongoing controversies surrounding bacterial species delimitation criteria[28–30]. By initiating classification at the genus level, we allow BacTaxID's hierarchical framework to resolve infrageneric diversity without imposing a priori species boundaries. Additionally, BacTaxID, like other k-mer-based distance estimation methods, exhibits reduced accuracy at genomic similarities below approximately 85% ANI. The relationship between ANI and Jaccard index follows the transformation $ANI = 1 - 1/k * \log(2J/(1+J))$[31], where k represents k-mer size and J denotes the Jaccard index. As shown in Supplementary Figure 1, this relationship maintains approximate linearity at Jaccard values exceeding 0.1, corresponding to ANI ≥ ~85%, but deviates substantially at lower similarities. This constraint necessitates establishing the genus as the minimum taxonomic level for reliable BacTaxID classification, as intergeneric comparisons frequently fall below the method's effective resolution threshold.

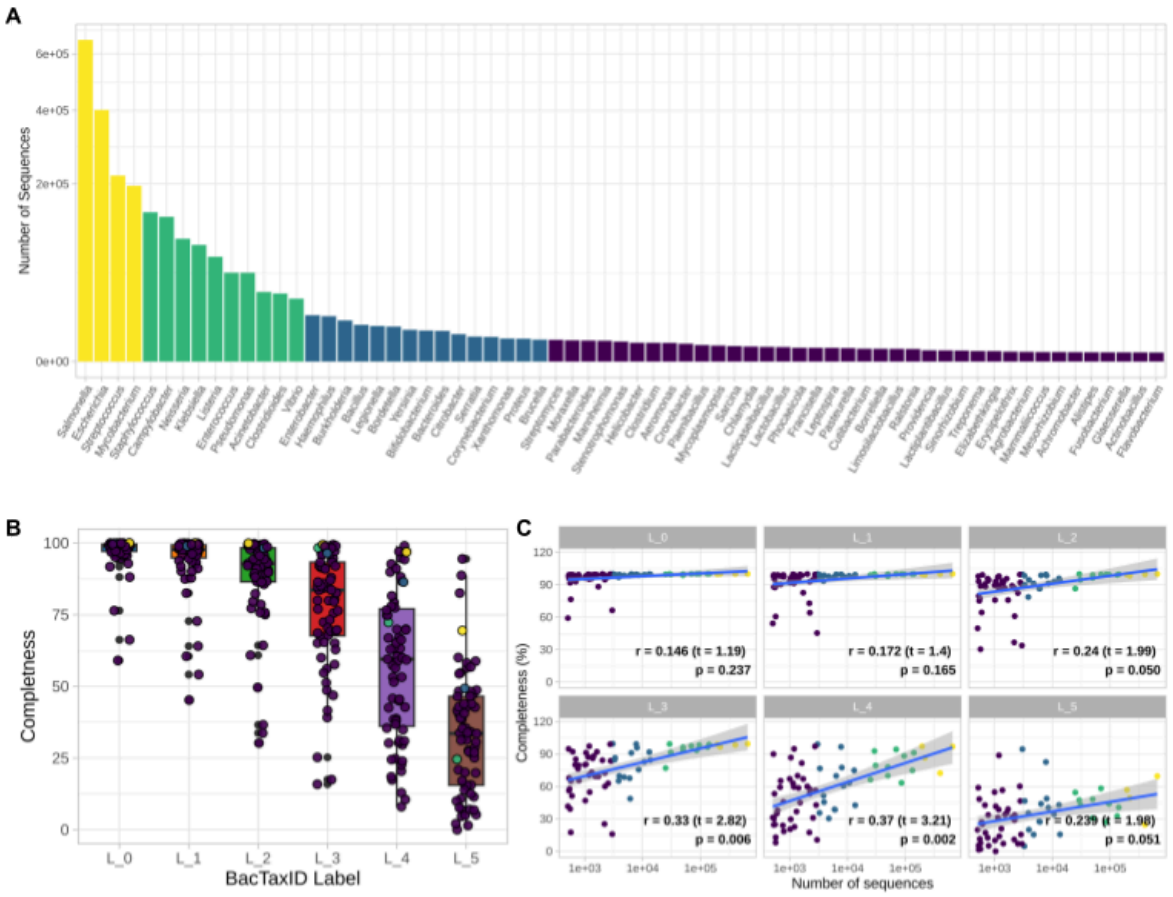


Figure 1. Dataset composition and typing scheme performance across bacterial genera. (A) Distribution of 2.3 million genomes across 67 genera from the All the Bacteria database, ordered by sample size. (B) Completeness distribution at each BacTaxID hierarchical level (L_0 through L_5, corresponding to ANI thresholds from 96% to 99.99%). Completeness indicates the percentage of genomes assigned to clusters at each resolution level. (C) Correlation between typing scheme completeness and genus-level dataset size across all six hierarchical levels. Each point represents a single genus, with colors denoting BacTaxID levels. Regression lines demonstrate positive correlation between sample size and scheme completeness, particularly at finer resolution levels. Pearson correlation coefficients (r), t-statistics (t), and p-values are displayed for each level.

Theoretically, the BacTaxID design improves classification performance as the dataset grows. The *click_size* parameter, which requires a minimum number of members to establish a new cluster, is intended to increase the completeness of the typing scheme by preventing the formation of spurious, low-density clusters in larger datasets. Here, we define completeness as the proportion of samples assigned a definitive code at each hierarchical level. Across the 2.3 million genomes analyzed, completeness exhibited high variability between levels, ranging from near-universal assignment (>98%) at the broader L_0 (96% ANI) resolution to significantly lower rates at the finest L_5 (99.99% ANI) level (Figure 1B). However, contrary to our initial hypothesis, our analysis revealed no consistent correlation between the number of genomes within a genus and the level-specific completeness of the BacTaxID scheme (Figure 1C). This finding suggests that the inherent population structure of the species or genus under study, rather than sheer dataset size, plays the determinant role in classification efficacy.

Multilevel Resolution of Bacterial Population Structure

To validate the biological relevance of our hierarchical scheme and assess concordance with established frameworks, we evaluated BacTaxID against consensus classifications in Escherichia and Salmonella (Figure 2). We computed Normalized Mutual Information (NMI)[32] between each hierarchical level and reference schemes to identify optimal alignment. Strong concordance emerged between species designations in the All the Bacteria database and L_0 clusters (96% ANI), reflecting the established 95% ANI species threshold[33]. However, some genomes classified as *Escherichia coli* and *Salmonella enterica*, the dominant species in each genus, clustered distinctly within different L_0 groups (Figure 2D, 2F), suggesting previously unrecognized or inadequately characterized pseudo-species not yet formally described.

For *Escherichia*, phylogroup assignments showed strong correlation with L_2 clusters (Figure 2D), indicating this resolution captures major phylogenetic subdivisions. At L_3, BacTaxID clusters demonstrated high concordance with MLST assignments, with NMI values indicating robust agreement between these independent methods (Figure 2C). This reflects the biological validity of fine-scale strain differentiation at this resolution, which aligns with epidemiologically relevant MLST resolution. Two additional levels (L_4 and L_5, corresponding to 99.5% and 99.99% ANI) capture sub-clonal diversity and micro-epidemiological variants beyond conventional MLST, providing superior discrimination for outbreak investigation and transmission tracking.

For Salmonella, we compared BacTaxID against SISTR[34] classifications (subspecies, serogroup, serovar, MLST, cgMLST). Similar to Escherichia, L_3 showed strong NMI correlation with serovar and MLST predictions (Figure 2E). Concordance with cgMLST increased progressively at finer resolutions, approaching maximum at L_5 (Figure 2E). However, agreement plateaued at L_4–L_5, revealing that the finest levels capture genomic variation beyond core genome content. This plateau indicates BacTaxID's sensitivity to whole-genome divergence rather than reliance on conserved loci alone.

The discordance between cgMLST and finest BacTaxID levels reflects differential weighting of accessory genome, plasmid content, and global sequence divergence, factors captured by sketch-based whole-genome comparison but excluded by core-genome typing. This positions BacTaxID as complementary to cgMLST, providing epidemiologically actionable discrimination while remaining sensitive to pangenomic heterogeneity. Collectively, these findings show BacTaxID generates biologically meaningful clusters that recapitulate established taxonomic relationships while providing finer-grained epidemiologically relevant resolution.

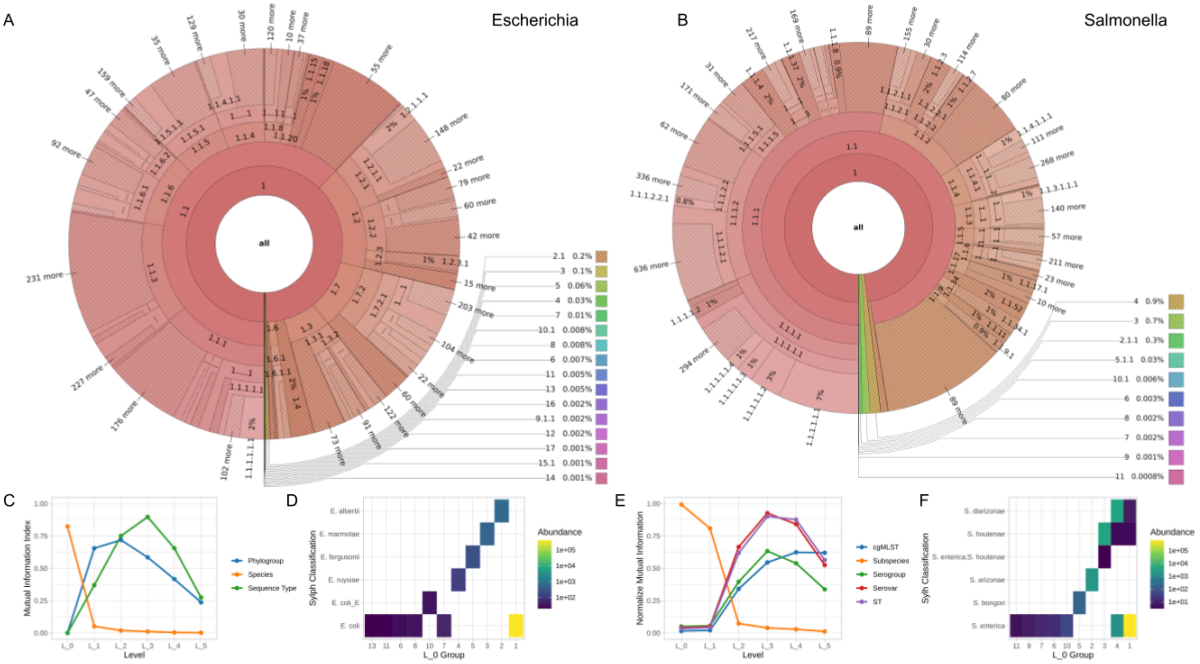


Figure 2. Population structure and agreement with established typing schemes in *Escherichia* and *Salmonella*. (A,B) Krona plots depicting the hierarchical population structure of *Escherichia* and *Salmonella* genera, respectively, as resolved by the BacTaxID typing scheme. Each concentric ring represents a successive hierarchical level (L_0 through L_5), with sector angles proportional to the relative abundance of genomes assigned to each cluster. Numerical labels indicate cluster identifiers, with annotations showing the number of additional clusters collapsed for visualization clarity. (C,E) Normalized Mutual Information (NMI) between BacTaxID hierarchical levels and established typing schemes for *Escherichia* and *Salmonella*, respectively. The x-axis denotes BacTaxID levels (L_0 through L_5), while the y-axis represents NMI values ranging from 0 (no agreement) to 1 (perfect concordance). Colored lines represent different reference typing systems: phylogroups, species, and sequence types for *Escherichia*; and cgMLST sequence types, subspecies, serovars, and serogroups for *Salmonella*. Higher NMI values indicate stronger correspondence between BacTaxID clusters and established classifications. (D,F) Species distribution within the first hierarchical level (L_0, 96% ANI threshold) for *Escherichia* and *Salmonella*, respectively. Tile plots display the number of genomes per species (y-axis, styled as classification) stratified by L_0 cluster assignment (x-axis). Color gradients represent genome abundance per species, illustrating the degree of species-level heterogeneity captured within coarse-resolution BacTaxID clusters.

To assess the generalizability of our findings beyond *Escherichia* and *Salmonella*, we extended our comparative analysis to 11 representative bacterial genera (*Acinetobacter, Enterobacter, Enterococcus, Escherichia, Haemophilus, Klebsiella, Neisseria, Pseudomonas, Salmonella, Staphylococcus,* and *Streptococcus*) comprising 1.7 million genomes, representing approximately 74% of the All the Bacteria collection (Figure 3). We evaluated BacTaxID concordance against both species designations and MLST classifications across these phylogenetically diverse lineages.

Consistent with observations from individual genera, species definitions exhibited high concordance with the L_0 and L_1 levels across all evaluated genera, demonstrating that BacTaxID's coarsest hierarchical resolutions reliably capture species-level genetic boundaries (Figure 3A). In contrast, MLST assignments showed a striking universal peak in concordance at L_3 (99% ANI threshold) across most genera (Figure 3B), with substantial variation in NMI trajectories at finer resolutions reflecting genus-specific differences in the relationship between pangenomic diversity and MLST genotype structure. This consistent optimization of MLST agreement at a single hierarchical level across phylogenetically independent bacterial lineages suggests that L_3 represents a universal epidemiological typing resolution, positioning BacTaxID as a genus-agnostic alternative to MLST that

maintains biological correspondence while capitalizing on whole-genome information rather than allelic profiles of housekeeping genes.

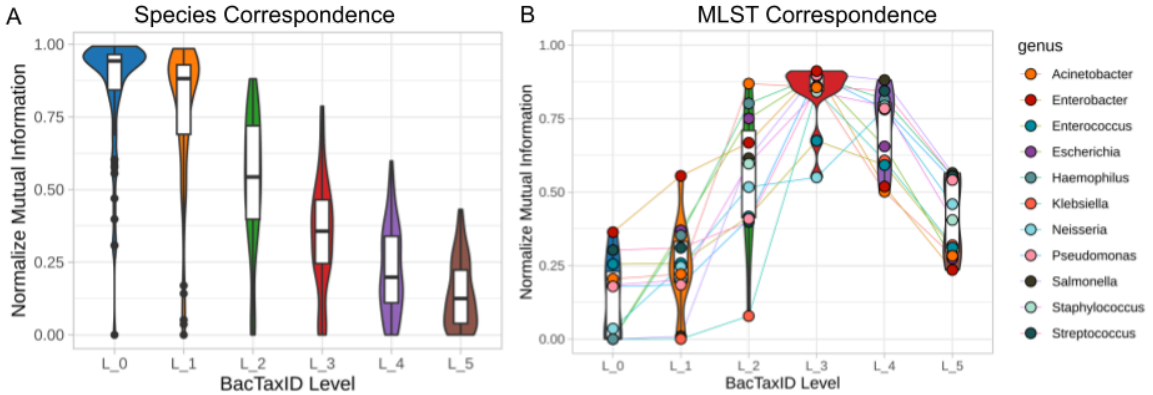


Figure 3. Consistency of BacTaxID typing performance across bacterial genera. (A) Normalized Mutual Information (NMI) between species designations and BacTaxID hierarchical levels across all 67 analyzed genera. Violin plots display the distribution of NMI values per level, showing high concordance at L_0 (96% ANI) and progressive decline at finer resolutions. (B) NMI between MLST assignments and BacTaxID levels across 11 representative bacterial genera. Individual genus trajectories demonstrate consistent peak concordance at L_3 (99% ANI threshold), supporting L_3 as a equivalent of universal MLST.

Hierarchical Resolution Optimization for Surveillance and Outbreak Investigation.

To demonstrate BacTaxID's utility across epidemiological contexts, we evaluated it in surveillance and outbreak detection scenarios, showcasing how adaptive resolution enables context-appropriate analysis from broad prevalence characterization to fine-scale transmission investigation.

For surveillance, we selected 30,000 E. coli genomes from All the Bacteria and stratified them by prevalence frequencies from a meta-analysis of extraintestinal pathogenic *E. coli* (ExPEC) infections[35], enabling direct comparison between MLST-based estimates and BacTaxID's hierarchical approach (Figure 4A, 4B).

At broadest resolution (L_1), a dominant cluster (1.2) comprised 52% of the population, encompassing STs 131, 12, 127, 73 and 95, followed by clusters 1.1 (22.3%) and 1.3 (17.3%) (Figure 4B). At intermediate L_2, cluster 1.2 subdivided into 1.2.1 (25.45%, almost entirely ST131) and 1.2.2 (32.34%, retaining other STs) (Figure 4C). At L_3, cluster distributions closely matched MLST prevalence (Figure 4D).

Critically, L_3 analysis revealed individual MLST types are not monophyletic. ST10 segregated across multiple L_3 clusters (Figure 4E), indicating substantial pangenomic diversity within a single designation. This polyphyly suggests MLST-based estimates conflate distinct sublineages, overestimating prevalence. BacTaxID's resolution enables more accurate strain characterization and risk stratification.

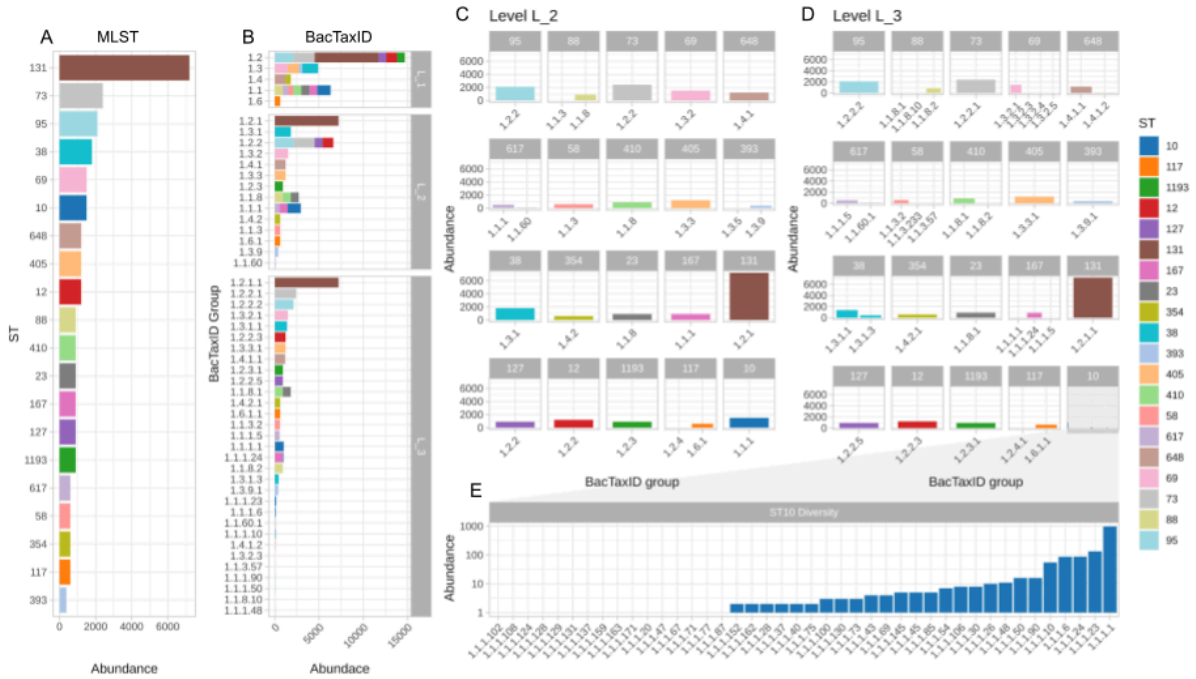


Figure 4. Multi-resolution strain classification in a prevalence survey context. (A) Distribution of MLST sequence types (ST) among *Escherichia coli* isolates from the All the Bacteria database stratified according to reported prevalence frequencies, ranked by abundance and colored by ST designation. (B) BacTaxID hierarchical clustering of the same isolate collection, displaying cluster assignments at levels L_1, L_2, and L_3 alongside their relative abundances. Only clusters containing more than 10 members are displayed for visualization clarity, demonstrating alternative resolution perspectives for the same genomic data. (C, D) Detailed distribution of BacTaxID hierarchical levels stratified by individual ST, comparing L_2 and L_3 resolutions. Bar plots display the abundance of each BacTaxID cluster nested within every detected ST, illustrating how intermediate hierarchical levels reveal sub-clonal structure not captured by MLST allelic profiling. (E) Fine-resolution structure within ST10, showing the distribution of 39 distinct lower-order BacTaxID groups at higher resolution levels, colored by ST assignment.

For outbreak detection, we reanalyzed two well-characterized cases. Lindemann *et al*[36]. documented a nosocomial *E. coli* ST38 (OXA-244) outbreak across three Norwegian hospitals (July–November 2020; 12 cases), with cgMLST showing 0–8 allelic differences within and 15–62 differences from unrelated ST38. Bojesen *et al*[37]. described seven broiler unit outbreaks over 18 months (June 2017–November 2018) from clonal E. coli ST10 (O132:H7) with ≤9 SNP differences among 36 isolates.

Outbreak definition depends on multifactorial assessment, but genomic similarity via cgMLST or SNP distances is central[38–40]. BacTaxID showed strong concordance with published definitions (Figure 5B–D). Computing SNP content among isolates in identical BacTaxID clusters across 10 clusters per species (Figure 5A): at L_4 (99.5% ANI), SNP density ranged from 41.5 SNPs/Mb (*Enterobacter*) to 416 SNPs/Mb (*Neisseria*); at L_5 (99.99% ANI), from 5.92 SNPs/Mb (*Acinetobacter*) to 27 SNPs/Mb (*Salmonella*). These show BacTaxID captures biologically meaningful variation, with L_5 offering sub-clonal resolution comparable to SNP analysis. Although L_5 may be insufficient for single-digit discrimination in the most stringent contexts, intermediate levels (L_3–L_4) enable early warning and rapid cluster identification in surveillance.

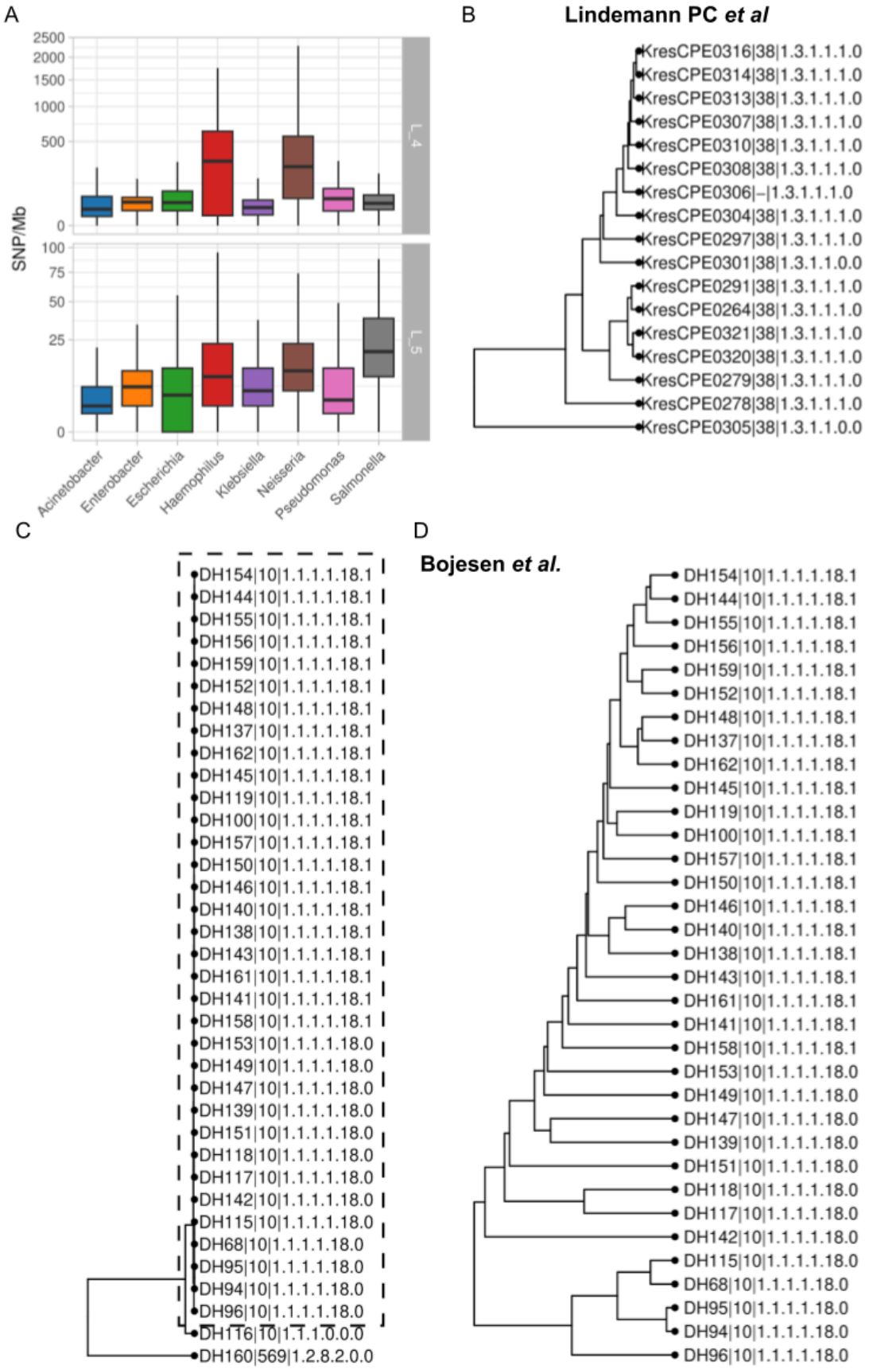


Figure 5. Outbreak detection capability and genomic equivalence of BacTaxID hierarchical levels. (A) Distribution of single nucleotide polymorphism (SNP) density per megabase within individual BacTaxID clusters at hierarchical levels L_4 (99.5% ANI) and L_5 (99.99% ANI) across 10 representative genera(B) Similarity tree of the Lindemann *et al. Escherichia coli* ST38 nosocomial outbreak, annotated with BacTaxID hierarchical codes at L_3. (C) Similarity tree of Bojesen *et al.* broiler outbreak of *Escherichia coli* ST10 poultry outbreak isolates, displaying the hierarchical classification. (D) Expanded similarity tree of the broiler outbreak isolates, illustrating fine-scale sub-clonal differentiation within the dominant outbreak cluster.

Discussion

Universality stems from whole-genome k-mer profiling rather than predefined loci, eliminating species-specific redesign and producing comparable outputs across clinical, evolutionary, and environmental domains. BacTaxID achieves standardized numeric classification through fully configurable parameters, sketch size controls sensitivity while similarity thresholds define specificity, enabling users to tailor resolution from species-level boundaries to strain-level discrimination. Unlike reference-dependent methods, this reference-free approach maintains strict proportionality to Average Nucleotide Identity (ANI), creating explicit quantitative links between vectorial distances and genomic divergence absent in traditional typing schemes. The resulting flexibility supports diverse applications: routine surveillance, outbreak investigation, and phylogeographic tracking across hierarchical levels that align with evolutionary boundaries (L_0 at 95% ANI, L_3 at 99% ANI concordant with MLST). Whole-genome operation thus delivers meaningful differentiation while transcending the limitations of predefined locus catalogs.

Retrospective analyses show BacTaxID complements, not replaces, established outbreak methods. In nosocomial E. coli ST38 and zoonotic E. coli ST10 outbreaks, it agreed with cgMLST or SNP-based definitions. L_5 (99.99% ANI) offers sub-clonal resolution comparable to SNP-based investigation (5.92–27 SNPs/Mb), but not single-digit SNP discrimination in the most stringent scenarios. Instead, the hierarchical design enables rapid prospective surveillance at L_3–L_4, providing early warning and guiding when to escalate to intensive molecular investigations.

Typing scheme completeness varies across levels and is not consistently correlated with dataset size. Completeness spans >98% at L_0 to often <80% at L_5, indicating that inherent population structure, evolutionary history, geography, ecology, matters more than sample count. For low-diversity or endemic pathogens, L_3–L_4 may be insufficient, whereas highly diverse organisms may be well-resolved at coarser levels, underscoring the need for context-specific validation.

A self-contained DuckDB format consolidates configuration, hierarchical identifiers, clustering metadata, and k-mer sketches into a single portable file, enabling fully autonomous genome profiling without external dependencies or centralized infrastructure, integrating via standard APIs (SQL, Python, R, CLI). This decentralized architecture reflects BacTaxID's semi-centralized design: local signature generation requires no centralized resources, yet standardized nomenclature across surveillance networks requires periodic inter-lab harmonization and centralized scheme distribution. The open-source, portable format facilitates adoption of independent profiling infrastructure while maintaining coordinated nomenclature consistency, balancing practical decentralization with standardization requirements.

For clinical deployment, integration with informatics systems and outbreak detection pipelines is advantageous. Resolution can adapt by context: L_2–L_3 for surveillance, L_4–L_5 for suspected outbreaks, with real-time escalation upon temporal clustering. The hierarchical

architecture also supports dynamic scheme evolution through local continuous clustering as databases grow, avoiding instability typical of flat approaches.

Pre-computed schemes for 67 genera provide a valuable community resource, though coverage is biased toward clinically relevant organisms, underrepresenting environmental, industrial, and understudied lineages. Community contributions of optimized schemes would broaden coverage and accelerate adoption. Overall, concordance with established frameworks, scalability to millions of genomes, and cross-taxon applicability position BacTaxID as a practical tool for surveillance, outbreak investigation, and microbial ecology, with ongoing work focused on population-structure effects and parameter optimization for diverse lineages.

Online Methods.

Sketch Generation and Distance Estimation

BacTaxID is implemented in Rust, a systems programming language that provides substantial computational advantages for large-scale bioinformatic workflows. Rust's memory safety guarantees eliminate entire classes of runtime errors including buffer overflows, use-after-free conditions, and data races, ensuring robust handling of diverse genomic datasets without sacrificing performance. Rust's native parallelization capabilities through packages like rayon allow seamless multi-threaded processing of large genome collections, enabling efficient utilization of modern multi-core processors. The resulting BacTaxID implementation achieves high throughput while maintaining strict resource management, making it suitable for processing millions of genomes.

BacTaxID employs a custom implementation combining binwise densified MinHash[23,24] with NtHash[25,26] for efficient k-mer hashing and genome sketching. The process consists of three integrated stages: genome parsing and preprocessing, k-mer extraction and hashing, and binwise signature construction.

Genome sequences are parsed from FASTA files (or compressed FASTA files) using the *needletail*[41] library, a high-performance Rust bioinformatics library that processes sequence records sequentially without loading entire files into memory. For each genome record, sequence normalization is applied to convert sequences to a canonical form (typically converting to uppercase and handling degenerate bases). The genome identifier (filename or sequence header) is extracted and normalized for integration into the hierarchical classification database. This stage establishes the foundation for consistent, reproducible sketch generation across diverse data sources.

Following normalization, each genome undergoes k-mer decomposition via a rolling-hash iterator implemented through the NtHash algorithm. Rather than explicitly generating and storing all k-mers (which would require N - k + 1 memory allocations per genome, where N is genome length), NtHash computes canonical k-mer hashes on-the-fly using a rolling window: as the window advances by one nucleotide, the hash is updated incrementally in constant time by removing the contribution of the departing nucleotide and incorporating the new incoming nucleotide. This streaming approach avoids materializing all k-mers, enabling memory-efficient processing even for large genomes (e.g., several Mb). The k-mer size parameter (denoted k) is configurable and defines the granularity of genomic information captured. Rust's ownership system and compile-time borrow checking guarantee memory safety throughout this iterative processing, preventing common programming errors while the compiler automatically optimizes the generated machine code for the target hardware. The resulting hash values are 64-bit unsigned integers uniformly distributed across the hash space [0, 2^64 - 1].

The binwise MinHash algorithm partitions the 64-bit hash space into a predefined number of bins (*sketch_size*, denoted S), where each bin spans approximately $\lceil 2^{64} / S \rceil$ contiguous hash values. As k-mer hashes are generated by the rolling iterator, each hash is assigned to its corresponding bin via integer division:

$$bin_{index} = \left\lceil \frac{hash}{bin_{size}} \right\rceil$$

Within each bin, only the minimum hash value encountered during sequence traversal is retained. This direct binning strategy operates in constant time per k-mer (O(1)), avoiding the O(log S) heap operations required by bottom-k MinHash schemes, while maintaining sketch quality comparable to traditional approaches. The resulting sketch is a fixed-length vector of S 64-bit minimum values (one per bin), effectively compressing the entire genome sequence into a compact representation.

The sketch vector is serialized to binary format using the *bincode*[42] library, which provides fast encoding and low-overhead representation without auxiliary metadata. This serialization enables rapid deserialization during hierarchical clustering operations and facilitates storage in the DuckDB[27] database. Duplicate k-mers within a sequence are automatically filtered through the streaming process: if the same k-mer appears multiple times, only the first occurrence contributes to the binwise construction, and subsequent duplicates do not affect the bin minimums, this behavior is inherent to the MinHash algorithm and requires no explicit duplicate removal step.

Pairwise genomic distances are computed by comparing sketch vectors: for two sketches S_1 and S_2, the number of matching bins (identical minimum hash values) is counted, and the Jaccard similarity is approximated as:

$$J = matches / S$$

where matches is the count of bins with identical values and S is the sketch size. This Jaccard estimate approximates the intersection-over-union of the underlying k-mer sets, capturing genomic overlap without explicit set operations. The relationship between Jaccard similarity and average nucleotide identity (ANI) is established through the Mash distance formula[31]:

$$D = 1/k \cdot \ln(2J/(1+J))$$

ANI is then calculated as ANI = 1 - D, and clamped to the range [0, 1]. This transformation maintains linearity and accuracy for ANI values ≥85% (corresponding to J ≥ 0.1), which encompasses the vast majority of meaningful bacterial comparisons; below this threshold, approximation error increases substantially, establishing the practical lower bound for BacTaxID's reliable operation.

Distance computations exploit merge-sort comparison of ordered hash arrays, achieving O(S) complexity per pairwise comparison. Parallelization via *rayon*[43] enables efficient all-against-all or hierarchical subset comparisons across millions of genomes, reducing wall-clock time to minutes even for massive datasets. In hierarchical searches (described in the Methods section on Hierarchical Clustering), comparisons are restricted to biologically relevant subsets, reducing effective complexity from $O(N^2)$ to O(N log N) or $O(N \sqrt{N})$.

All sketch data, including serialized binwise signatures, genome identifiers, and hierarchical classification results, are consolidated into a self-contained DuckDB database file. This unified format eliminates external dependencies on separate sketch files and ensures full reproducibility: any researcher can query results decades later without relying on external services or archived databases. The database supports standard SQL queries as well as interfaces via Python, R, and command-line tools, enabling seamless integration into diverse analytical pipelines. Sketch parameters (k, S) are stored alongside results, ensuring future replicability and enabling cross-validation of findings.

Hierarchical Clustering and Code Assignment

BacTaxID implements a recursive hierarchical clustering strategy that processes genomes level-by-level, progressively assigning hierarchical classification codes while maintaining biological consistency and computational efficiency. The algorithm operates through two primary pathways: direct assignment to existing clusters (classifier search) or creation of new clusters via graph-theoretic clique detection (de novo group formation).

The system processes genomes through a sequence of hierarchical levels (L_0, L_1, ..., L_N), each with a predefined distance threshold corresponding to a specific ANI boundary (e.g., L_0 = 96% ANI, L_3 = 99% ANI). Thresholds are stored in the database as ordered pairs (level_name, distance_threshold) and retrieved during initialization. The algorithm traverses levels in ascending order of stringency: coarser resolutions (lower ANI thresholds, earlier levels) are evaluated first, and finer resolutions (higher ANI thresholds, later levels) only when necessary. This order exploits the hierarchical constraint that a genome's code at level i must be consistent with its code at level i-1: specifically, if a genome is assigned to group 1.3 at L_1, then its L_0 code must be 1 (the parent group). This monotonicity constraint is enforced at every assignment step and prevents inconsistent classifications.

Within each cluster at a given level, genomes receive one of two designation states: classifier (C) or satellite (S). Classifier genomes serve as references for subsequent hierarchical searches and new isolate assignments. Satellite genomes are assigned to clusters based on similarity to classifiers but do not themselves serve as references, this distinction prevents spurious clustering caused by outliers, recombinants, or hypermutants that may satisfy distance thresholds but represent atypical lineages. The decision to assign classifier status depends on two parameters: *click_threshold* (typically 80%) and *reference_size* (typically 100 genomes per cluster). Specifically, a candidate genome is promoted to classifier status if: (1) its distance to a reference cluster is below the level threshold, (2) the fraction of reference cluster members within the distance threshold is at least *click_threshold*, and (3) the current cluster size remains below *reference_size*. This criterion prevents indefinite cluster expansion while permitting controlled growth as new genomes with substantial similarity arrive.

When a new query genome is being classified at level i, the algorithm first retrieves all classifier genomes from the database that share the query's parent-level code (e.g., if the query was assigned to group 1.3 at L_1, only classifiers with L_1 code = 1.3 are retrieved). The query sketch is then compared against these reference sketches to compute pairwise distances. If any reference sketch satisfies the level-specific distance threshold, the query is

assigned to that cluster, inheriting the cluster's hierarchical code. The closest reference (best hit) is prioritized, but consistency checks ensure that the reference cluster's code maintains the required parent-level prefix. If a valid best hit exists and classifier criteria are met, the query is marked as a classifier at that level. This pathway is computationally efficient because the comparison space is restricted to biologically coherent subsets rather than all genomes, achieving O(log N) or O(√N) complexity per level compared to O(N) for flat clustering schemes.

If no valid best hit is identified at a given level (or if the best hit has code = 0 indicating an unassigned reference), the algorithm transitions to graph-theoretic clique detection. In this pathway, the algorithm retrieves all genomes (both classifiers and satellites) that share the query's parent-level code, computes all pairwise distances between the query and these references, and constructs a distance graph where nodes represent genomes and edges connect pairs whose distance falls below the level threshold. The system then searches for maximal cliques, complete subgraphs where all pairwise distances satisfy the threshold, using iterative deepening and constraint propagation. A clique is considered valid for group formation if its size meets the minimum *click_size* threshold (typically 5 genomes). Upon discovery of a valid clique, the query is assigned a new unique code at that level (the next unused integer for that parent group), and all clique members are updated with this new classification. Importantly, non-clique members in the distance graph receive the same new code but with state = S (satellite), allowing them to be associated with the emerging cluster without influencing future reference searches.

Hierarchical codes are constructed incrementally as the algorithm descends through levels. At L_0, codes are simple integers (1, 2, 3, ...). At subsequent levels, codes are concatenated with dots: if a genome is assigned to group 5 at L_0, group 3 at L_1, and group 7 at L_2, its full hierarchical code is 5.3.7. This hierarchical identifier uniquely specifies the genome's position in the similarity tree and enables efficient database queries via string prefix matching. The *code_full* field is maintained alongside the numeric code value for both human readability and computational efficiency in hierarchical filtering operations.

The algorithm processes each level sequentially and terminates at a given level under three conditions: (1) a valid classifier-state assignment is achieved and the genome is classified as a classifier, (2) the genome has been assigned via clique detection and new cliques no longer form at finer levels, or (3) no valid distances or cliques are detected at a particular level, leaving that level (and all finer levels) unassigned. This design permits incomplete classifications, where a genome may be fully classified up to L_3 but lack assignments at L_4 and beyond. Incompleteness is expected and acceptable when population structure does not support finer-scale clustering, and provides an explicit signal to downstream analyses that higher-resolution typing is not possible or not meaningful for that particular genome within its biological context.

All parameters governing hierarchical classification (number of levels, level-specific thresholds, *click_size*, *click_threshold*, *reference_size*) are user-configurable and stored in the database metadata table upon scheme initialization. This flexibility allows genus-specific or species-specific tuning: organisms with higher evolutionary rates may require coarser thresholds at given ANI boundaries to capture meaningful epidemiological groups, whereas organisms with lower diversity may achieve adequate resolution at finer thresholds. The

algorithm scales linearly with the number of query genomes (each processed independently via parallel workflows) and logarithmically with database size per level (due to hierarchical space partitioning), enabling efficient processing of millions of genomes with commodity hardware.

Classification results are stored in a centralized database table that maintains hierarchical information for each genome: the numeric classification code at each level (e.g., 5, 3, 7), the complete hierarchical designation as a concatenated string (e.g., 5.3.7), and the genomic role (classifier or satellite). Database operations employ transactional integrity protocols to ensure that if an error occurs during the writing of classification assignments, the entire operation is rolled back rather than corrupting the database with incomplete or inconsistent data. This prevents cascading errors and maintains data reliability even if processing is interrupted. Additionally, an optional debug mode logs all intermediate computational steps, including genome-to-reference distances and clique detection results, to facilitate retrospective validation of classification decisions and troubleshooting of ambiguous or unexpected assignments.

The update command orchestrates the complete workflow: it loads the SketchManager from the DuckDB database, processes each input FASTA file to generate a query sketch, executes hierarchical classification through iterative level traversal, and persists the resulting sketch and classification codes to the database. This modular design enables efficient incremental updates: new genomes can be added to an existing scheme without re-clustering previously classified samples, maintaining deterministic identifier assignments and enabling long-term stability of nomenclature across laboratory networks.

Graph construction and clique detection

BacTaxID employs a graph-theoretic framework for de novo cluster formation when a query genome cannot be assigned to an existing group. This process involves dynamic graph construction from pairwise distance data and the application of maximal clique detection algorithms to identify internally consistent groups of genomes.

When direct assignment fails at a given hierarchical level, the algorithm constructs an undirected graph in memory to represent pairwise genomic relationships. First, it retrieves all genomes that share the same parent-level code as the query genome but are not yet classified at the current level (i.e., their hierarchical code for this level is empty). This set includes the query genome itself along with any previously processed genomes that failed to be assigned to a cluster at this resolution. Pairwise distances between all genomes in this unclassified set are computed, and this distance data is inserted as edges into a temporary edges table in the DuckDB database.

The in-memory graph is then built using the *petgraph*[44] library. Each unique genome identifier is added as a node, and edges are established between nodes if their pairwise distance exceeds the ANI threshold for the current hierarchical level. This is achieved by querying the edges table with dynamic SQL, filtering for node pairs present in the current reference set and whose distance satisfies the level threshold.

Once the graph is constructed, BacTaxID searches for maximal cliques, subgraphs where every node is connected to every other node, using the *maximal_cliques* algorithm from the petgraph library. A clique represents a set of genomes that are all mutually similar above the specified distance threshold, ensuring high internal consistency and preventing the chaining artifacts common in single-linkage or agglomerative clustering methods.

The algorithm iterates through all identified maximal cliques and evaluates them against the *click_size* parameter (typically 5 genomes). Because the classification process is dynamic and genomes are added sequentially, it is highly unusual for more than one new clique to form simultaneously. The introduction of a single new query genome is typically insufficient to bridge two previously separate, unclassified groups into distinct, valid cliques at the same time. Therefore, the algorithm considers the first clique found that meets or exceeds the minimum size as the valid new cluster. This "first-found" heuristic balances computational cost with biological representativeness, as larger cliques typically represent well-established, densely populated regions of the genomic landscape, and the sequential nature of the algorithm makes the discovery of multiple concurrent cliques a rare event.

Upon identification of a valid clique, a new hierarchical code is generated for that level. The system queries the code table to determine the next available integer code for the current parent group (e.g., if the highest existing code for parent 1.3 is 1.3.5, the new clique is assigned code 1.3.6). This new code is assigned to the query genome, and its status is set to classifier (C).

Crucially, this new classification is then propagated to all other members of the identified clique. Each genome within the clique has its hierarchical code updated in the database to reflect membership in the new cluster. Other genomes present in the initial distance graph but not part of the validated clique are also assigned the new code, but with a satellite (S) status. This ensures they are associated with the newly formed group but do not act as references in future classification steps.

To optimize subsequent hierarchical searches, the algorithm performs edge pruning. After a clique is processed at level j, all edges that constituted the clique are deleted from the edges table if their distance is below the threshold for the next finer level (j+1). This prevents redundant re-evaluation of these high-similarity relationships in subsequent, more stringent clique searches. This strategic pruning ensures that the edges table remains lean and that clique detection at finer resolutions operates on a reduced, relevant subset of the data, further enhancing computational scalability.

Creation of the All the Bacteria Classification Scheme

BacTaxID classification schemes were generated for 67 bacterial and archaeal genera selected from the All the Bacteria[22]all (ATB) database based on the availability of ≥500 quality-controlled genomes per genus. The ATB collection comprised 2.4 million genome assemblies as of August 2024, yielding a total of 2.3 million genomes distributed across 67 genera and 3,926 species for hierarchical classification. Individual typing schemes were constructed for each genus using uniform parameterization (*sketch_size* = 4000, *kmer_size* = 31, *click_size* = 5, *click_threshold* = 0.8, *reference_size* = 100), with six hierarchical

clustering levels corresponding to average nucleotide identity (ANI) thresholds of 96%, 98%, 99%, 99.5%, 99.9%, and 99.99%. Each genome was assigned a hierarchical typing code composed of up to six integers separated by periods (for example, 1.3.1.8.12.1), where each integer represents the cluster assignment at successive levels of resolution. All pre-computed schemes were deposited in a public repository and are freely accessible with accompanying metadata, classification results, and interactive exploration tools at www.bactaxid.org and https://doi.org/10.5281/zenodo.17226627.

Statistical Analysis

Correlation between genome number within each genus and level-specific classification completeness was assessed using pearson correlation coefficient. Completeness was defined as the proportion of genomes assigned definitive hierarchical codes at each successive level of resolution.

Biological validity of the hierarchical framework was systematically evaluated by computing Normalized Mutual Information (NMI)[32] between BacTaxID classifications at each hierarchical level and independent reference typing schemes, including MLST, core-genome MLST (cgMLST), and species designations. Initial validation analyses were performed on two model organisms, *Escherichia* and *Salmonella*, and subsequently extended across 11 representative genera (*Acinetobacter, Enterobacter, Enterococcus, Escherichia, Haemophilus, Klebsiella, Neisseria, Pseudomonas, Salmonella, Staphylococcus,* and *Streptococcus*), encompassing 1.7 million genomes representing approximately 74% of the All the Bacteria collection. NMI values were computed to quantify the degree of concordance between classification systems, with values ranging from 0 (independent) to 1 (perfect agreement), enabling direct comparison of clustering quality across hierarchical levels.

To further assess the discriminatory power and epidemiological relevance of BacTaxID classifications, a prevalence study simulation was conducted using *Escherichia coli* as a representative model organism. A random subset of 30,000 whole-genome sequences of E. coli was selected from the All the Bacteria collection and subjected to BacTaxID classification across all hierarchical levels.

To assess epidemiological utility, comprehensive outbreak detection analyses were performed on well-characterized nosocomial and zoonotic *Escherichia coli* transmission events with documented epidemiological linkage and independent molecular confirmation. For each outbreak, BacTaxID cluster assignments at all hierarchical levels were compared against published clinical and epidemiological definitions derived from core-genome MLST (cgMLST) profiles and pairwise single nucleotide polymorphism (SNP) distance matrices. Phylogenetic relationships among outbreak-associated isolates were inferred using UPGMA (Unweighted Pair Group Method with Arithmetic Mean)[45] hierarchical clustering based on BacTaxID k-mer distance metrics, providing visual assessment of clonal structure and reconstruction of probable transmission chains.

Within-cluster SNP density was quantified by calculating pairwise SNP differences among isolates assigned to identical BacTaxID codes at hierarchical levels L_4 (99.5% ANI) and L_5 (99.99% ANI) across eight representative genera (*Acinetobacter, Enterobacter, Escherichia,*

Haemophilus, Klebsiella, Neisseria, Pseudomonas, and *Salmonella*). SNPs between genome pairs were identified using the *snps_pairwise* function from pato[46], which performs individual pairwise genome alignments to identify single nucleotide polymorphisms and calculate SNP density normalized to genome alignment. Results were expressed as SNPs per megabase (SNPs/Mb) of whole genome alignment, enabling standardized comparison across genera with varying genome sizes and evolutionary rates.

Statistical Software and Visualization

All descriptive statistical analyses were conducted using R (version 4.5)[47] with the following complementary packages: *tidyverse*[48] for data manipulation and visualization pipelines, *DBI*[49] and *duckdb*[50] for efficient database query and integration, *ggsci*[51] for scientific color palettes, *patchwork*[52] for multi-panel figure composition, *ggpubr*[53] for publication-ready graphical outputs, *tidyinftheo*[54,55] for information-theoretic calculations, *ggrepel*[56] for optimized text label placement, *ggnewscale*[57] for multiple color scales within single plots, and *ggtree*[58,59] for phylogenetic tree visualization and annotation.

Additionally, comprehensive interactive hierarchical classification visualizations were generated for all 68 bacterial and archaeal genera using KronaTools[60], a specialized bioinformatics tool designed for multilevel taxonomic and metagenomic data visualization (available at https://github.com/marbl/Krona). KronaPlots enable rapid exploration of hierarchical BacTaxID classifications and provide intuitive assessment of genus-level and strain-level diversity across the complete taxonomy. All genus-specific KronaPlots have been integrated and are publicly accessible through the BacTaxID web platform at www.bactaxid.org, facilitating interactive data exploration and dynamic visualization of hierarchical classification schemes for researchers and public health practitioners.

Genus-Specific Supplementary Typing Annotations

To provide enhanced epidemiological and evolutionary context, standardized typing annotations were obtained for select genera using specialized in silico typing resources:

Salmonella: Supplementary serovar, serotype, cgMLST and MLST were obtained through SISTR[34] (Salmonella In Silico Typing Resource), which performs rapid in silico serotyping and subspecies assignment based on genomic sequences.

Klebsiella and *Neisseria*: Annotation data including strain typing and virulence factors were sourced from the All the Bacteria supplementary resource, available at https://allthebacteria.readthedocs.io/en/latest/typing.html.

Escherichia: Phylogroups were determined in silico using ClermontTyping[61], a specialized classification tool designed for rapid phylogenetic assignment of *Escherichia* isolates into recognized phylogroups (A, B1, B2, D, E, F, and cryptic), accessible at https://github.com/A-BN/ClermonTyping.

Multilocus Sequence Typing (MLST) for bacterial genera not subjected to specialized typing analysis, conventional MLST sequence typing was performed using the standardized

pipeline available at https://github.com/tseemann/mlst. This tool performs allelic profiling against appropriate species-specific MLST schemes and assigns sequence types (STs) according to established PubMLST[62,63] nomenclature.

Software availability and web resources

Pre-computed hierarchical classification schemes for all 68 bacterial and archaeal genera are publicly accessible without restrictions through the BacTaxID web platform (www.bactaxid.org), with full classification data, metadata, and interactive visualization tools provided at no cost. Complete datasets, including genomic identifiers, hierarchical codes, and supplementary annotations for all 2.3 million genomes in the All the Bacteria collection, have been deposited in the Zenodo research repository (https://doi.org/10.5281/zenodo.17226627) to ensure long-term preservation and accessibility.

The BacTaxID software framework, including source code and precompiled binaries, is available through open-source distribution at https://github.com/irycisBioinfo/BacTaxID under the GPL-3.0 license. Reproducible analysis scripts, complete data processing workflows, statistical analyses, and supplementary metadata tables supporting the results presented in this study are publicly available at https://github.com/irycisBioinfo/BacTaxID_publication_analysis/.

Acknowledgements and Funding

V.F.-L. is supported by the “Miguel Servet program” from the ISCIII to promote professional research careers in biomedicine and health sciences in NHS centers (CP22/00164). MDFdB was supported by the Instituto de Salud Carlos III (pFIS F19/00366) and V.F.L Miguel Servet funding (CP22/00164).

References

1. Sabat, A. J. *et al.* Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **18**, 20380 (2013).
2. Simar, S. R., Hanson, B. M. & Arias, C. A. Techniques in bacterial strain typing: past, present, and future. *Curr. Opin. Infect. Dis.* **34**, 339–345 (2021).
3. Shropshire, W. C., Hanson, B. M. & Shelburne, S. A. Genome-wide approaches to bacterial strain typing: a history and review of recent methodological advances. *Curr. Opin. Infect. Dis.* **38**, 329–338 (2025).
4. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 3140–3145 (1998).
5. Uelze, L. *et al.* Typing methods based on whole genome sequencing data. *One Health Outlook* **2**, 3 (2020).
6. Zhong, L. *et al.* Distributed genotyping and clustering of Neisseria strains reveal continual emergence of epidemic meningococcus over a century. *Nat. Commun.* **14**, 7706 (2023).
7. Khdhiri, M. *et al.* refMLST: reference-based multilocus sequence typing enables universal bacterial typing. *BMC Bioinformatics* **25**, 280 (2024).
8. Maiden, M. C. J. *et al.* MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **11**, 728–736 (2013).
9. Li, H. *et al.* KleTy: integrated typing scheme for core genome and plasmids reveals repeated emergence of multi-drug resistant epidemic lineages in Klebsiella worldwide. *Genome Med.* **16**, 130 (2024).
10. Bleidorn, C. & Gerth, M. A critical re-evaluation of multilocus sequence typing (MLST) efforts in Wolbachia. *FEMS Microbiol. Ecol.* **94**, (2018).

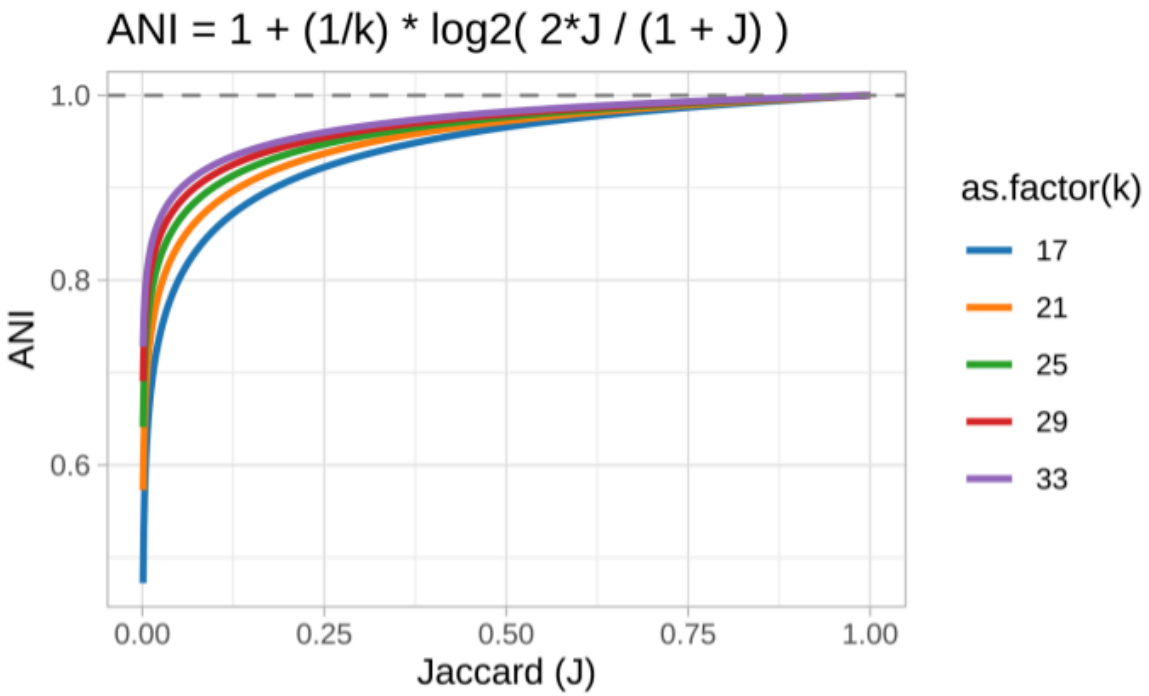
11. Blanc, D. S., Magalhães, B., Koenig, I., Senn, L. & Grandbastien, B. Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of Pseudomonas aeruginosa. *Front. Microbiol.* **11**, (2020).
12. Miro, E. *et al.* Core/Whole Genome Multilocus Sequence Typing and Core Genome SNP-Based Typing of OXA-48-Producing Klebsiella pneumoniae Clinical Isolates From Spain. *Front. Microbiol.* **10**, (2020).
13. Zhou, Z., Charlesworth, J. & Achtman, M. HierCC: a multi-level clustering scheme for population assignments based on core genome MLST. *Bioinformatics* **37**, 3645–3646 (2021).
14. Tian, L., Huang, C., Mazloom, R., Heath, L. S. & Vinatzer, B. A. LINbase: a web server for genome-based identification of prokaryotes as members of crowdsourced taxa. *Nucleic Acids Res.* **48**, W529–W537 (2020).
15. Marakeby, H. *et al.* A System to Automatically Classify and Name Any Individual Genome-Sequenced Organism Independently of Current Biological Classification and Nomenclature. *PLOS ONE* **9**, e89142 (2014).
16. Hennart, M. *et al.* A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of Klebsiella pneumoniae Strains. *Mol. Biol. Evol.* **39**, msac135 (2022).
17. Duchêne, S. *et al.* Genome-scale rates of evolutionary change in bacteria. *Microb. Genomics* **2**, e000094 (2016).
18. Rocha, E. P. C. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Mol. Biol. Evol.* **35**, 1338–1347 (2018).
19. Torrance, E. L., Burton, C., Diop, A. & Bobay, L.-M. Evolution of homologous recombination rates across bacteria. *Proc. Natl. Acad. Sci.* **121**, e2316302121 (2024).
20. van Belkum, A. *et al.* Guidelines for the validation and application of typing methods for use in bacterial epidemiology. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* **13 Suppl 3**, 1–46 (2007).

21. Bien, J. & Tibshirani, R. Hierarchical Clustering With Prototypes via Minimax Linkage. *J. Am. Stat. Assoc.* **106**, 1075–1084 (2011).
22. Hunt, M. *et al.* AllTheBacteria – all bacterial genomes assembled, available, and searchable. 2024.03.08.584059 Preprint at https://doi.org/10.1101/2024.03.08.584059 (2025).
23. Zhao, J., Zhao, X., Pierre-Both, J. & Konstantinidis, K. T. BinDash 2.0: New MinHash Scheme Allows Ultra-fast and Accurate Genome Search and Comparisons. 2024.03.13.584875 Preprint at https://doi.org/10.1101/2024.03.13.584875 (2024).
24. Zhao, X. BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinforma. Oxf. Engl.* **35**, 671–673 (2019).
25. Mohamadi, H., Chu, J., Vandervalk, B. P. & Birol, I. ntHash: recursive nucleotide hashing. *Bioinformatics* **32**, 3492–3494 (2016).
26. Kazemi, P. *et al.* ntHash2: recursive spaced seed hashing for nucleotide sequences. *Bioinformatics* **38**, 4812–4813 (2022).
27. Raasveldt, M. & Mühleisen, H. DuckDB: an Embeddable Analytical Database. in *Proceedings of the 2019 International Conference on Management of Data* 1981–1984 (Association for Computing Machinery, New York, NY, USA, 2019). doi:10.1145/3299869.3320212.
28. Simar, S. R., Hanson, B. M. & Arias, C. A. Techniques in bacterial strain typing: past, present, and future. *Curr. Opin. Infect. Dis.* **34**, 339–345 (2021).
29. Rosselló-Móra, R. & Amann, R. Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.* **38**, 209–216 (2015).
30. Barco, R. A. *et al.* A Genus Definition for Bacteria and Archaea Based on a Standard Genome Relatedness Index. *mBio* **11**, e02475-19 (2020).
31. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
32. Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. Comparing community structure identification. *J. Stat. Mech. Theory Exp.* **2005**, P09008 (2005).

33. Rodriguez-R, L. M. *et al.* An ANI gap within bacterial species that advances the definitions of intra-species units. *mBio* **15**, e02696-23 (2023).
34. Yoshida, C. E. *et al.* The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. *PLOS ONE* **11**, e0147101 (2016).
35. Manges, A. R. *et al.* Global Extraintestinal Pathogenic Escherichia coli (ExPEC) Lineages. *Clin. Microbiol. Rev.* **32**, 10.1128/cmr.00135-18 (2019).
36. Lindemann, P. C. *et al.* Intraregional hospital outbreak of OXA-244-producing Escherichia coli ST38 in Norway, 2020. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **28**, 2200773 (2023).
37. Bojesen, A. M., Ahmed, U., Skaarup, H. & Espinosa-Gongora, C. Recurring outbreaks by the same Escherichia coli ST10 clone in a broiler unit during 18 months. *Vet. Res.* **53**, 2 (2022).
38. Payne, M. *et al.* Enhancing genomics-based outbreak detection of endemic Salmonella enterica serovar Typhimurium using dynamic thresholds. *Microb. Genomics* **7**, 000310 (2021).
39. Carson, J., Keeling, M., Ribeca, P. & Didelot, X. Incorporating Epidemiological Data into the Genomic Analysis of Partially Sampled Infectious Disease Outbreaks. *Mol. Biol. Evol.* **42**, msaf083 (2025).
40. Duval, A., Opatowski, L. & Brisse, S. Defining genomic epidemiology thresholds for common-source bacterial outbreaks: a modelling study. *Lancet Microbe* **4**, e349–e357 (2023).
41. needletail - crates.io: Rust Package Registry. https://crates.io/crates/needletail (2025).
42. bincode - crates.io: Rust Package Registry. https://crates.io/crates/bincode (2025).
43. rayon - crates.io: Rust Package Registry. https://crates.io/crates/rayon (2025).
44. petgraph - crates.io: Rust Package Registry. https://crates.io/crates/petgraph/0.8.1 (2025).

45. R.r, S. & C.d, M. A STATISTICAL METHOD FOR EVALUATING SYSTEMATIC RELATIONSHIPS. **38**, 1409–1438 (1958).
46. Fernández-de-Bobadilla, M. D. *et al.* PATO: Pangenome Analysis Toolkit. *Bioinformatics* **37**, 4564–4566 (2021).
47. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2025).
48. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
49. R Special Interest Group on Databases (R-SIG-DB), Wickham, H. & Müller, K. *DBI: R Database Interface*. (2024). doi:10.32614/CRAN.package.DBI.
50. Mühleisen, H. & Raasveldt, M. *Duckdb: DBI Package for the DuckDB Database Management System*. (2025).
51. Xiao, N. *Ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'Ggplot2'*. (2025).
52. Pedersen, T. L. *Patchwork: The Composer of Plots*. (2025).
53. Kassambara, A. *Ggpubr: 'ggplot2' Based Publication Ready Plots*. (2025).
54. Pohl, A. *Tidyinftheo: Some Information-Theoretic Functions in the 'Tidy' Style*. (2025).
55. Cover, T. M. *Elements of Information Theory*. (John Wiley & Sons, 1999).
56. Slowikowski, K. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. (2024). doi:10.32614/CRAN.package.ggrepel.
57. Campitelli, E. *Ggnewscale: Multiple Fill and Colour Scales in 'Ggplot2'*. (2025). doi:10.32614/CRAN.package.ggnewscale.
58. Yu, G. *Data Integration, Manipulation and Visualization of Phylogenetic Treess*. (Chapman and Hall/CRC, 2022). doi:10.1201/9781003279242.
59. Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
60. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

61. 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. Genomics* **4**, e000192 (2018).
62. 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.* **3**, 124 (2018).
63. Jolley, K. A. & Maiden, M. C. J. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).



Supplementary Figure 1. Theoretical relationship between the Jaccard index (J) and the Average Nucleotide Identity (ANI). The plot illustrates the Mash-derived equation for estimating ANI from J, a k-mer based similarity metric. Each colored line represents the estimated ANI value for a given Jaccard index, calculated using different k-mer sizes (k), ranging from 17 to 33. The dashed line indicates the theoretical maximum of 1.0 for both ANI and J.