ASM | Applied and Environmental Microbiology

| Environmental Microbiology | Full-Length Text

Whole-genome surveillance supports hazard profiling of *Escherichia coli* lineages in recycled water treatment systems

Dewa A. P. Rasmika Dewi,[1] Harshanie Abeywardena,[2] Dieter Bulach,[3] Chi-Wen Tseng,[4] Lamiya Bata,[4] Fiona Lynch,[1] Melita Stevens,[2] Nicholas D. Crosbie,[2] Rebekah Henry[1,4]

AUTHOR AFFILIATIONS See affiliation list on p. 15.

ABSTRACT The use of treated wastewater is increasingly important for sustainable water management under a changing climate, yet conventional monitoring based on *Escherichia coli* enumeration provides limited insight into strain diversity and associated public health hazards. Here, we applied longitudinal whole-genome sequencing (WGS) to 180 *E. coli* isolates collected across the treatment continuum of a recycled water facility, from influent to final effluent. Genomic analysis revealed extensive strain-level heterogeneity, comprising 88 sequence types across eight phylogroups, with greater diversity in influent than in treated effluent. Phylogenetic comparisons with contextual Australian genomes indicated clustering with strains associated with companion animals, wild birds, humans, and livestock, suggesting multiple potential source reservoirs rather than a single dominant origin, although source contributions were not definitive. Despite a >90% reduction in total *E. coli* loads, isolates recovered from upstream and downstream stages exhibited broadly comparable virulence factor and antimicrobial resistance gene (ARG) profiles, suggesting that, within the cultured isolate collection, reductions in abundance exceeded shifts in genomic composition. To assess operational relevance, we prototyped a genomics-informed hazard framework integrating virulence determinants, ARGs, plasmid-associated mobility, and reuse-specific exposure context. Using this framework, 92.8% of isolates were classified as low hazard, and 7.2% as moderate hazard, with no isolates meeting criteria for high or critical hazard classifications. These findings demonstrate that genomic profiling of indicator organisms can reveal population structure and hazard heterogeneity not captured by conventional enumeration alone, and can provide a practical basis for incorporating genomic information into hazard-informed monitoring of recycled water systems.

IMPORTANCE Routine recycled water monitoring relies largely on culture-based *E. coli* counts, which indicate regulatory compliance but provide limited insight into strain diversity, persistence, and genomic characteristics relevant to public health. Using longitudinal whole-genome sequencing, we show that genetically distinct *E. coli* lineages, including isolates carrying combinations of virulence and antimicrobial resistance determinants, can persist through advanced treatment despite substantial reductions in overall *E. coli* loads. While most isolates were classified as low genomic hazard and no high- or critical-hazard isolates were detected, these findings demonstrate that conventional enumeration alone cannot distinguish between genetically diverse lineages with differing hazard potential in highly treated systems. By integrating genomic data into a hazard classification framework, this study demonstrates an applied approach to contextualize *E. coli* detections and distinguish low-risk background populations from isolates with elevated genomic hazard profiles. This work supports the use of genomic profiling of indicator organisms to improve surveillance, inform treatment performance assessment, and enable more risk-based management of recycled water systems.

Editor Satoshi Ishii, University of Minnesota Twin Cities, St. Paul, Minnesota, USA

Address correspondence to Dewa A. P. Rasmika Dewi, mika.dewi@monash.edu.

The authors declare no conflict of interest.

See the funding table on p. 16.

Received 12 July 2026
Accepted 25 August 2026
Published 15 September 2026

Copyright © 2026 Dewi et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

KEYWORDS recycled water, *Escherichia coli*, whole genome sequencing, antimicrobial resistance, virulence factors, microbial hazard framework, water reuse

Increasing global water demand, driven by climate variability and population growth, has accelerated the adoption of wastewater recycling as a key component of sustainable water management (1). Recycled water is now widely used in agricultural irrigation, industry, environmental augmentation, and, in some regions, indirect or direct potable reuse (2). As reliance on water reuse increases, ensuring microbial safety across diverse end uses remains a critical operational and regulatory challenge.

Escherichia coli is widely used as a regulatory indicator of fecal contamination and treatment performance. However, this species exhibits substantial genomic diversity, encompassing environmentally persistent commensals as well as lineages carrying virulence factors (VFs) and antimicrobial resistance genes (ARGs) of clinical significance (3, 4). Consequently, equivalent culture-based counts may represent markedly different public health risks. Although ultraviolet (UV) irradiation and chlorination effectively reduce microbial loads (5), viable *E. coli* can still be detected following treatment. Evidence from environmental and engineered systems indicates that some strains can persist in water, sediments, or soils and tolerate fluctuating environmental conditions (6–8). In recycled water schemes, such persistence raises important questions regarding lineage-specific resilience and the distribution of clinically relevant genetic traits.

Whole-genome sequencing (WGS) enables high-resolution characterization of *E. coli* lineage structure, accessory genome diversity, and the presence of virulence determinants, ARGs, and mobile genetic elements (MGEs) (9, 10). While genomic approaches have been applied to natural aquatic environments (11, 12), their integration into routine assessment of recycled water systems remains limited (13). Engineered treatment systems provide a controlled context in which inputs, processes, and operational conditions are defined, enabling evaluation of how treatment influences microbial population structure. Genomic analysis in these systems supports surveillance beyond binary compliance metrics and enables more resolved assessment of microbial hazard.

Integrating genomic data with operational metadata, including treatment stage, temporal variation, and end-use exposure context, can improve hazard assessment and support evidence-based decision-making. Identification of persistent lineages and clinically relevant traits provides actionable information for optimizing treatment performance and managing risk across reuse applications.

This study applied longitudinal whole-genome sequencing to characterize lineage diversity, persistence across treatment stages, and the distribution of VFs, ARGs, and plasmids in an engineered recycled water system. These data were integrated with system-level and end-use information to evaluate a genomics-informed hazard framework for surveillance and decision-making.

RESULTS

Spatial-temporal dynamics and environmental drivers of *E. coli* loads

To examine spatial variation within the recycled water system, *E. coli* loads (MPN/100 mL) were compared across sampling sites and seasons. Significant spatial variation was observed, with the highest *E. coli* concentrations in influent samples and substantially lower concentrations following treatment. Water samples from W1 exhibited the highest loads, with a median $\log_{10}$(MPN + 1) of 1.82 (IQR: 1.57–2.07), corresponding to 64.5 MPN/100 mL. In comparison, W3 had a median of 0.30 (IQR: 0.18–0.64), and W2 showed the lowest levels (median = 0, IQR: 0–0.18) (Table S1A). Wilcoxon rank-sum tests confirmed significant differences between all sources: W1 vs W2 (W = 1119, $P = 8.18 \times 10^{-13}$), W1 vs W3 (W = 2869, $P < 2.2 \times 10^{-16}$), and W3 vs W2 (W = 520, $P = 1.45 \times 10^{-7}$) (Fig. S1A).

Temporal patterns were also examined, and correlation between overall *E. coli* load and physicochemical parameters (pH and temperature) were assessed. *E. coli* loads were higher in winter (median = 1.91; IQR: 0.35–2.14) than in other seasons (spring: 0.35;

summer: 0.48; autumn: 0.48). However, the overall season effect was not significant (χ^2 = 6.20, *df* = 3, *P* = 0.102) (Table S1B, Fig. S1B). Dunn's post-hoc tests showed no significant pairwise differences. The spring–winter comparison approached significance (*P* = 0.091), whereas all other comparisons were not significant (Table S1C).

To explore environmental drivers, *E. coli* concentration data ($\log_{10}$(MPN + 1)) were correlated with physicochemical parameters at W1 and W3. No physicochemical data were available for W2; therefore, these samples were excluded. Aggregated data from W1 and W3 showed a significant positive correlation with pH (Spearman's ρ = 0.41, *P* < 0.001; Fig. S1C) and a weak negative correlation with temperature (ρ = –0.26, *P* = 0.040; Fig. S1D). However, these associations were not observed when analyzed by site (pH: W1 ρ = –0.31, *P* = 0.092; W3 ρ = –0.12, *P* = 0.520; temperature: W1 *P* = 0.158; W3 *P* = 0.687). For dissolved oxygen (DO), no association was observed in aggregated data (ρ = –0.10, *P* = 0.456; Fig. S1E), whereas a moderate negative correlation was detected at W1 (ρ = –0.41, *P* = 0.024).

Genomic diversity of *E. coli* isolates and potential source attribution

All 180 genomes passed quality control thresholds (>90% completeness and <5% contamination). Sequencing coverage ranged from 21× to 393× (median = 45×). Sequencing coverage ranged from 21× to 393× (median = 45×). Assemblies comprised 1–25 contigs per genome, with N50 values ranging from 0.28 to 5.22 Mb (median = 1.46 Mb) (Table S2). These assembly metrics indicate high-quality draft genomes suitable for downstream comparative genomic and phylogenetic analyses. Whole-genome sequencing of 180 *E. coli* isolates (W1: *n* = 85; W2: *n* = 11; W3: *n* = 84) revealed high genomic heterogeneity, encompassing 88 distinct sequence types (STs). The most common STs were ST155 and ST162 (each *n* = 8; 4.4%) (Table S2). Twenty-five isolates (13.9%) possessed allelic profiles that did not match existing sequence types in EnteroBase and were therefore classified as untyped STs. Requests for assignment of new ST designations have been submitted and are currently under review by the EnteroBase curators.

Isolates from W1 exhibited greater ST richness (56 unique STs) compared to those from W3 (42 unique STs). Phylogroup distribution spanned eight groups, with B1 predominating (*n* = 120 isolates, 66.7%), followed by A (*n* = 23, 12.8%), B2 (*n* = 11, 6.1%), D (*n* = 11, 6.1%), and E (*n* = 9, 5%). Less common phylogroups included C (*n* = 3; 1.7%), G (*n* = 1; 0.6%), and H (*n* = 1; 0.6%), and one isolate was unassigned (unknown, *n* = 1; 0.6%) (Table S2).

The total gene content per genome ranged from 4,073 to 5,067 genes (Table S3). Mean gene counts were similar between W1 (4,440 ± 136) and W3 (4,521 ± 191), with comparable ranges across sites (Table S3, Fig. S2). Due to the limited number of isolates recovered from W2 (*n* = 11) and their restricted temporal coverage (October 2024–January 2025 and May 2025), W2 was excluded from formal statistical comparisons to avoid bias associated with unequal sample sizes and sampling effort. Gene richness differed significantly between W1 and W3 (Kruskal–Wallis χ^2 = 6.53, df = 1, *P* = 0.011), although the absolute difference was modest. PERMANOVA based on Jaccard distances indicated a small but significant shift in gene composition between sites (R^2 = 0.016, $F_{1,167}$ = 2.65, *P* = 0.002). Homogeneity of dispersion was not significant ($F_{1,167}$ = 2.01, *P* = 0.159).

Phylogenetic analysis of the 180 study isolates together with 121 closely related Australian reference genomes (ANI ≥99% to at least one study isolate) is shown in Fig. 1. The majority of isolates (*n* = 68, 37.8%) clustered most closely with companion animal-associated contextual genomes, primarily canines, followed by wild birds (*n* = 44, 24.4%), humans (*n* = 25, 13.9%), and livestock (pigs: 23 isolates, 12.8%; cattle: six isolates, 3.3%) (Table S4, Fig. S3A). Smaller proportions were similar to poultry-associated strains (*n* = 12, 6.7%), wastewater (*n* = 1, 0.6%), and feline sources (*n* = 1, 0.6%). Source distributions showed evidence of variation between W1 and W3 (Fisher's Exact test, *P* = 0.027), although these patterns should be interpreted cautiously given the sampling design and isolate numbers. While companion animal associations were similar at both sites (~40%),

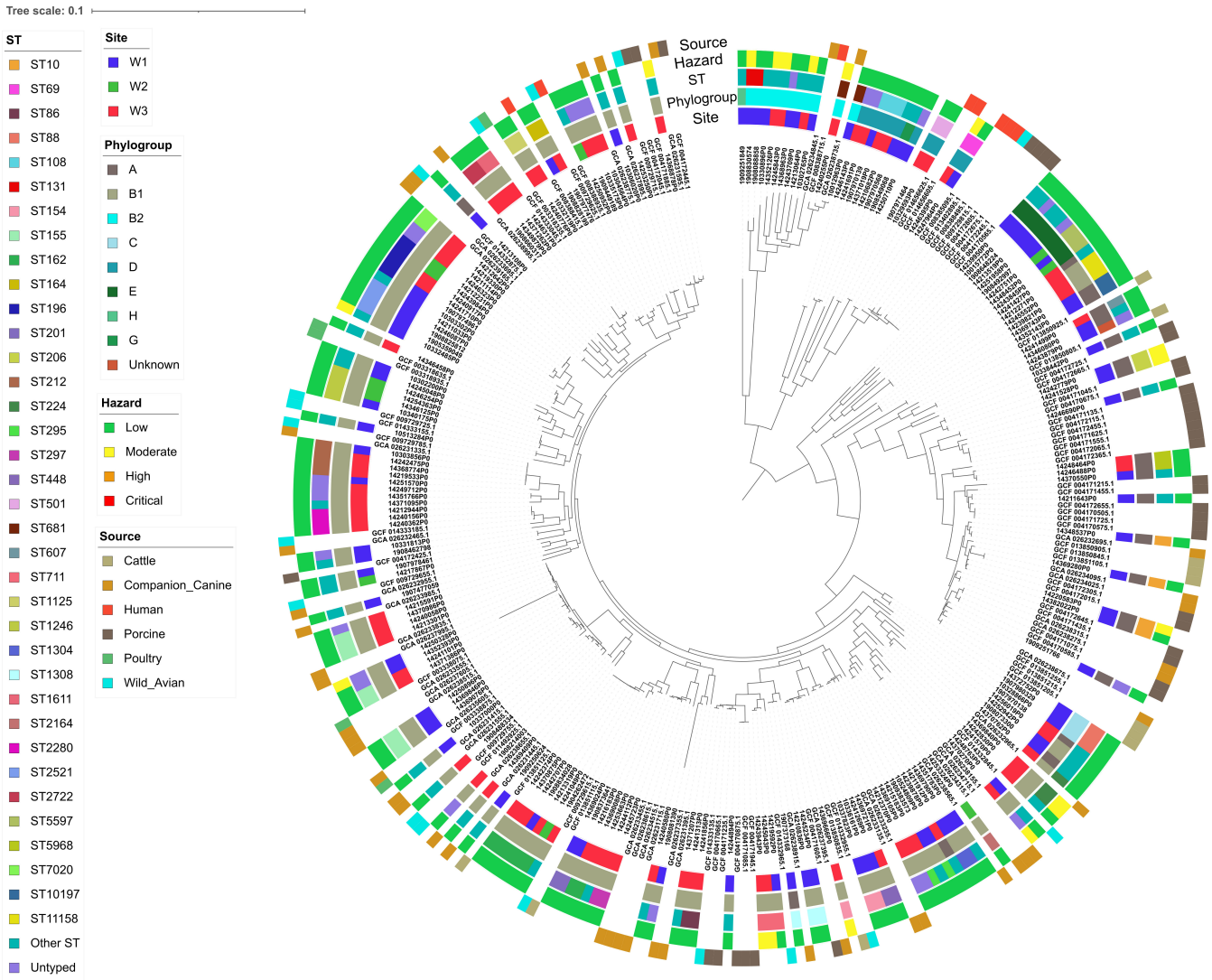


FIG 1 Core genome phylogenetic tree of 180 *E. coli* isolates from the recycled water system and 121 contextual *E. coli* genomes from historical Australian collections sharing ≥99% average nucleotide identity (ANI). The tree was constructed from a core-genome SNP alignment using IQ-TREE v2.1.3 with 1,000 bootstrap replicates and visualized using iTOL (https://itol.embl.de/). Metadata tracks adjacent to each tip indicate sampling location, phylogroup, sequence type (ST), genomic hazard classification, and the associated isolation source of the Australian reference genomes (companion animal, livestock, wild animal, human, and wastewater). For visualization purposes, STs represented by more than two isolates were assigned individual colors in the ST legend, whereas STs detected in two or fewer isolates were grouped and displayed as "Other ST." Untyped isolates were displayed separately as "Untyped." The tree was midpoint rooted.

secondary associations differed, with W1 more frequently linked to porcine sources and W3 to wild birds (Fig. S3B). These findings suggest that *E. coli* populations within the recycled water system likely originate from multiple host reservoirs rather than a single dominant source. The increased representation of wild bird-associated lineages among W3 isolates may reflect persistence through treatment, ecological selection within the post-treatment environment, residence within the contact basin or associated infrastructure, or other post-treatment environmental processes.

Composition and population structure of *E. coli* isolates

Phylogroup compositions differed among sites (Fisher's exact test, $P = 0.010$), with rare phylogroups (C, G, and H; $n < 5$) grouped as "Other." Although statistically significant, these results should be interpreted cautiously because unequal isolate recovery across sampling locations may have influenced the observed distributions. Both W1 and W3

were dominated by phylogroup B1, but W1 showed greater heterogeneity, whereas W3 lacked several minor phylogroups (Fig. 2A). Correspondingly, α-diversity was higher at W1 than W3 (Shannon: 1.37 vs 0.80; Simpson: 0.65 vs 0.38).

Sequence type distributions also suggested differences across sites (Fisher's exact test, $P = 2 \times 10^{-4}$; W1 vs W3, $P = 0.003$). Diversity indices indicated slightly higher richness and evenness at W1 (Shannon 3.85; Simpson 0.975) than at W3 (3.66; 0.970). Several STs were shared between sites, including ST162, ST155, and ST212 (Fig. 2B), suggesting the possibility of persistence or repeated introduction of certain lineages across treatment stages. However, confirmation of these patterns would benefit from a larger sample size to improve statistical confidence. SNP analysis of the most prevalent STs (ST162: $n = 7$ and ST155: $n = 8$) revealed contrasting patterns. ST162 included near-identical genomes (≤1 SNP) across sites and timepoints, as well as highly divergent pairs (>3,000 SNPs), indicating co-occurrence of closely related and distinct lineages (Table S5A). In contrast, ST155 isolates showed substantial divergence, with only two closely related intra-site pairs (<10 SNPs) and no closely related inter-site pairs (Table S5B). These findings suggested that ST155 populations were largely composed of highly diverse lineages with only limited evidence of local clonal expansion.

Virulence determinants and antibiotic resistance profiles

A total of 195 distinct virulence genes were detected across all isolates (Table S6). The number of unique genes detected per site differed (W1: 181; W3: 127; W2: 68), although this likely reflects differences in sample size. The median number of virulence genes per genome was identical across sites (median = 46), with similar variability (Fig. 3A). No significant difference was observed between W1 and W3 ($W = 4114$, $P = 0.082$). Presence–absence analysis of virulence gene profiles revealed only minor differences between the W1 and W3 isolates. Although PERMANOVA detected a statistically significant effect of collection site on virulence composition ($R^2 = 0.013$, $P = 0.041$), site explained only 1.3% of the total variation, and dispersion analysis indicated unequal within-site variability (betadisper $P = 0.024$), suggesting broadly conserved virulence repertoires across treatment stages.

A total of 47 distinct ARGs were identified across the isolate collection (Table S6). Site-level ARG richness was highest in W1 (39 ARGs) and W3 (34 ARGs) and lowest in W2 (seven ARGs). Median ARG counts per genome were similar across sites (~5 genes per genome), with no significant difference between W1 and W3 (Wilcoxon rank-sum test: W = 3596, $P = 0.928$) (Fig. 3B). The dominant resistance-associated profiles

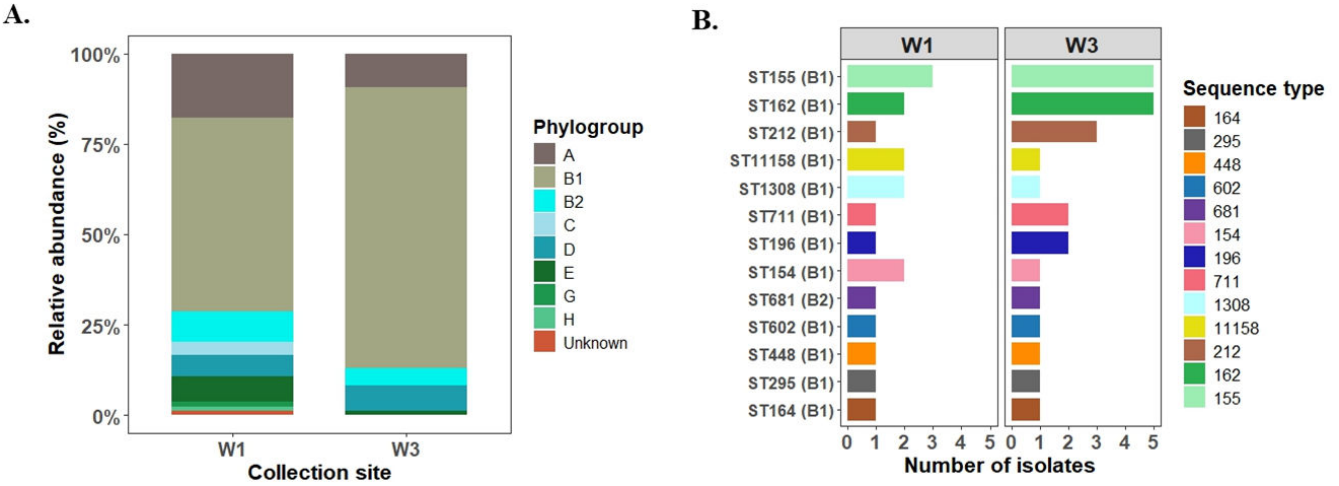


FIG 2 (A) Proportional distribution of *E. coli* phylogroups in influent (W1) and post-treatment (W3) samples. (B) Distribution of the 10 most prevalent sequence types detected in both influent (W1) and post-treatment (W3) samples, excluding untyped isolates (NT). Sequence types are ordered by overall frequency, and their corresponding phylogroups are indicated alongside each ST label.

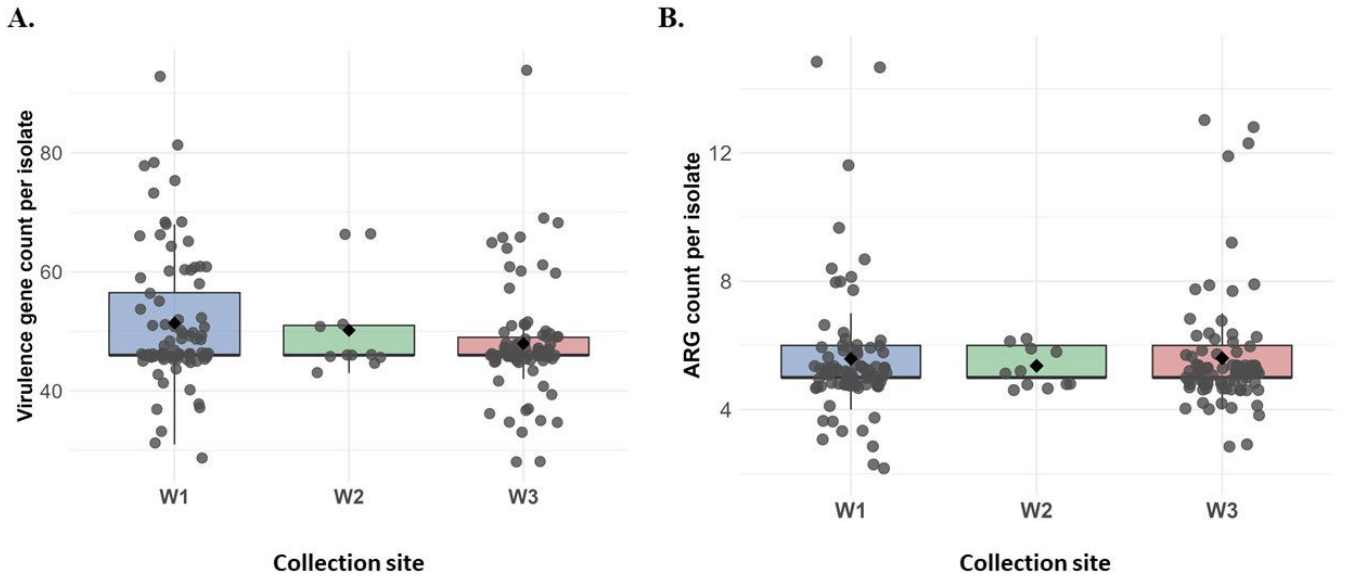


FIG 3 Richness of virulence factors (VFs) and antibiotic resistance genes (ARGs) in *Escherichia coli* isolates across collection sites (W1, W2, W3). (A) Boxplots show the distribution of VF counts per isolate at each site. Boxes represent the interquartile range, the horizontal line indicates the median, and gray dots show individual isolates. The black diamond marks the mean VF count for each site. Colors correspond to collection sites. (B) Boxplots of the total number of ARG counts per isolate at each site, irrespective of whether individual resistance determinants represent intrinsic resistance mechanisms in *E. coli*. The same visual conventions as in panel A are used: interquartile range (boxes), median (line), individual isolates (gray dots), and mean (black diamond).

were broadly similar between influent and post-treatment isolates. Efflux-associated genes and β-lactam determinants were detected in almost all isolates from both W1 and W3, while fosfomycin-associated genes occurred in more than 90% of isolates at both sites (Fig. S4). Colistin-associated determinants were more prevalent in W3 but were almost entirely represented by chromosomal *pmrB* variants (*pmrB_Y358N* and *pmrB_E123D*) rather than acquired *mcr* genes. Other resistance subclasses, including quinolone, sulfonamide, tetracycline, and cephalosporin determinants, occurred at low frequencies and exhibited similar prevalence across treatment stages (Fig. S4). Overall, W1 and W3 exhibited broadly similar resistance profiles despite differences in isolate abundance across sites.

Genomic hazard assessment framework

The genomic hazard assessment framework integrated virulence determinants, antimicrobial resistance genes, and plasmid-associated mobility into a composite hazard classification interpreted within the context of recycled water end use (EAHI) (Table S7). Overall, most isolates were classified as low hazard (92.8%, n = 167), while a smaller proportion were classified as moderate hazard (7.2%, n = 13) (Table S7; Fig. 4A). No isolates met the criteria for high or critical hazard classifications.

Hazard profiles varied across sampling locations (Fig. 4B). Influent samples (W1) were predominantly composed of low-hazard isolates (98.8%, n = 84), with one isolate classified as moderate hazard (1.2%). All isolates recovered from the intermediate treatment stage (W2) were classified as low hazard (n = 11). Post-treatment samples (W3) contained a higher proportion of moderate-hazard isolates (14.3%, n = 12), although low-hazard isolates remained predominant (85.7%, n = 72). The hazard-profile distribution differed significantly between W1 and W3, with lower odds of moderate versus low hazard at W1 than at W3 (Fisher's exact test, P = 0.0012; OR = 0.07, 95% CI = 0.002–0.51).

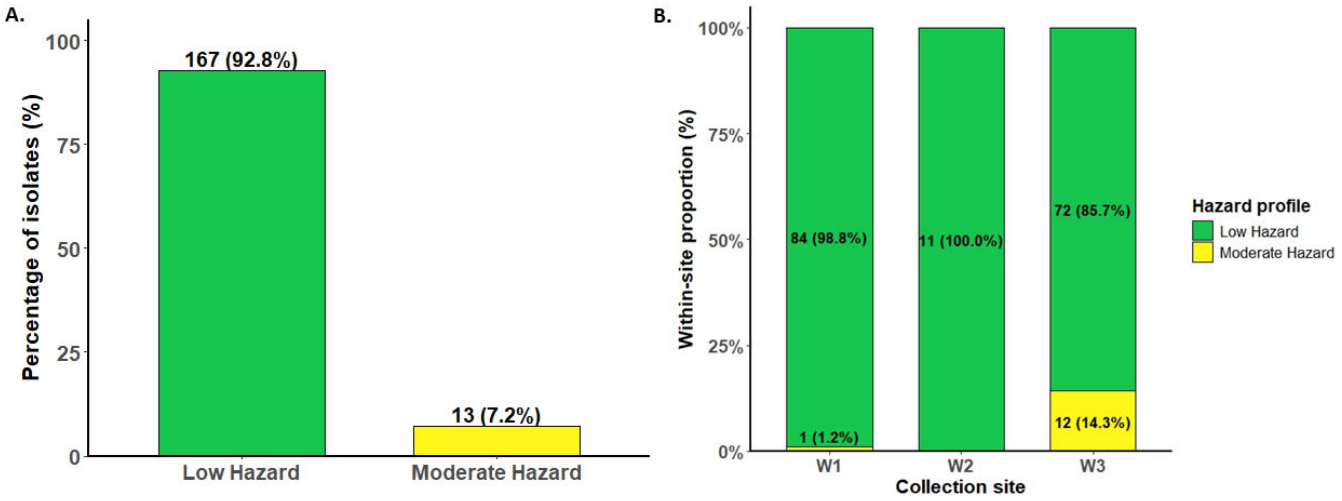


FIG 4 Genomic hazard profile of *E. coli* isolates recovered from the recycled water system. (A) Overall hazard distribution of isolates classified as low (green), moderate (yellow), high (orange), or critical (red) according to the genomic hazard assessment framework. Bars represent the percentage of isolates assigned to each hazard category, with labels indicating the number and proportion of isolates (*n*, %). (B) Hazard profile across sampling sites (W1, W2, and W3). Stacked bars show the within-site proportion of isolates assigned to each hazard category, with labels indicating both the number of isolates and the corresponding percentage within each site (*n*, %).

DISCUSSION

Routine microbial monitoring of recycled water has traditionally relied on culture-based enumeration of *E. coli* to assess treatment performance and regulatory compliance (14, 15). While cost-effective and operationally useful, this approach provides limited resolution of strain diversity and ecological behavior. In contrast, genomic profiling provides higher-resolution insight into population structure, persistence, and the distribution of virulence and resistance determinants in engineered water systems. This improved resolution enables differentiation between low-risk background detections and isolates carrying genomic features associated with elevated hazard potential (9, 10).

Dynamics of *E. coli* load and environmental drivers

Spatial patterns confirmed substantial reduction in *E. coli* loads along the treatment train, with highest concentrations at W1 and lowest at downstream sites. Although winter medians were higher, seasonal differences were not statistically significant, suggesting that short-term variability rather than consistent seasonal effects likely explains observed fluctuations (16, 17).

Correlations between *E. coli* loads and physicochemical parameters were observed in aggregated data but not within individual sites. This indicates that these relationships likely reflect differences between treatment stages rather than within-site environmental effects. A moderate negative association between dissolved oxygen and *E. coli* at W1 may reflect localized ecological conditions; however, effect sizes were modest and should be interpreted cautiously. Overall, these findings suggest that treatment processes are the primary determinants of *E. coli* abundance, with physicochemical variation playing a secondary role within the observed operational range.

Genomic diversity and source attribution

Whole-genome sequencing revealed substantial genomic heterogeneity, encompassing 88 sequence types across multiple phylogroups. The predominance of B1 and A lineages is consistent with environmentally adapted populations commonly observed in aquatic systems (18–20), whereas the presence of B2 and D phylogroups indicates occasional occurrence of lineages associated with higher virulence potential (21, 22). Reduced

phylogroup and ST diversity at W3 relative to W1 suggests selective filtering along the treatment continuum, with downstream dominance of environmentally resilient lineages.

Phylogenetic comparisons indicated similarity to strains associated with companion animals, wild birds, humans, and livestock, suggesting that the *E. coli* population within the recycled water system likely reflects multiple fecal reservoirs rather than a single dominant source. Companion animal-associated lineages were the most frequently identified, consistent with increasing recognition of pets as reservoirs of environmentally disseminated *E. coli*, while the substantial contribution of wild bird-associated lineages may reflect interactions between recycled water infrastructure and surrounding wildlife. Differences in inferred source distributions between W1 and W3 may reflect treatment-associated selection, differences in environmental exposure following treatment, or limitations inherent to reference-based source attribution approaches.

Accumulation of organic material may support localized persistence or regrowth of *E. coli* (23, 24). Importantly, source attribution based on ≥99% ANI should be interpreted cautiously, as similarity to reference genomes reflects phylogenetic relatedness and ecological association rather than definitive host origin, particularly for generalist B1 lineages (25, 26). Nevertheless, these findings highlight the utility of whole-genome sequencing for providing source-related context that is not obtainable through conventional *E. coli* enumeration. Such information may assist water utilities in identifying likely contamination pathways, evaluating treatment performance, and prioritizing management interventions aimed at reducing microbial hazards within recycled water systems.

Although differences in gene composition between W1 and W3 were statistically significant, the effect size was small (PERMANOVA R^2 = 1.6%). Therefore, these findings

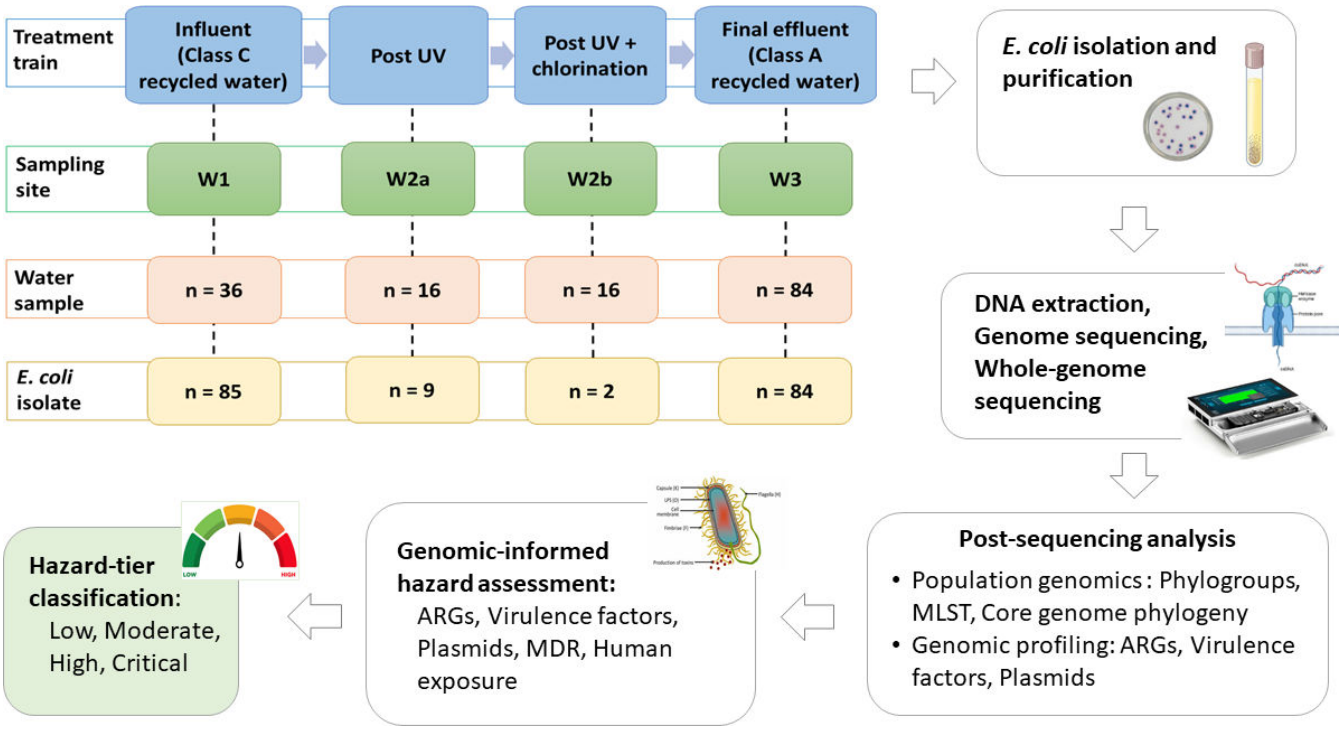


FIG 5 Overview of the study design and analytical workflow. Water samples were collected from influent Class C recycled water (W1), post-UV treatment (W2a), post-UV plus chlorination treatment (W2b), and final Class A recycled water stored in the contact basin prior to distribution (W3). Sampling yielded 180 *E. coli* isolates, which were subjected to culture purification, Oxford Nanopore whole-genome sequencing, genome assembly, taxonomic validation, phylogenetic and population genomic analyses, and characterization of virulence factors, antimicrobial resistance genes (ARGs), and plasmid-associated mobile genetic elements. Results were integrated into a genomic hazard framework incorporating virulence potential, ARG hazard, multidrug resistance breadth, plasmid hazard, and potential human exposure pathways.

should be interpreted as modest shifts in population structure rather than major changes in the overall genomic composition of *E. coli* across the treatment continuum. Further studies incorporating larger and more balanced isolate collections would help clarify the extent to which these patterns are representative of *E. coli* populations within recycled water systems.

Lineage dynamics and persistence

SNP analysis revealed lineage-specific dynamics within dominant sequence types. ST162 included both near-identical genomes across sites and highly divergent lineages, indicating co-occurrence of closely related and distinct strains within the same MLST group. The presence of near-identical isolates across sites and timepoints is consistent with short-term persistence or repeated introduction, whereas highly divergent pairs likely reflect independent inputs or longer-term diversification.

In contrast, ST155 isolates showed substantial genomic divergence, with limited evidence of close relatedness and no closely related inter-site pairs. This pattern is consistent with repeated introduction rather than sustained persistence within the treatment system. These findings highlight variability in lineage behavior and demonstrate the value of genome-scale analysis for distinguishing persistence from reintroduction in engineered water systems (23, 27).

Virulence and antimicrobial resistance profiles

Across all isolates, 195 virulence genes and 47 ARGs were detected. Despite differences in site-level gene richness, the number of virulence genes per genome was consistent across sites, indicating that treatment reduced overall diversity without substantially altering per-genome virulence content. Many detected virulence determinants were associated with adhesion, iron acquisition, secretion systems, and colonization functions that may contribute to environmental persistence and host interaction (28, 29). This pattern is consistent with selection for traits that enhance survival under nutrient-limited and intermittently disinfected conditions (30–32).

ARG abundance per genome did not differ significantly between sites, suggesting that treatment influenced lineage composition without substantially changing the resistance gene burden of surviving populations. Efflux-associated genes were ubiquitous across isolates; however, their prevalence should be interpreted cautiously because many represent intrinsic components of the *E. coli* genome that contribute to baseline resistance phenotypes rather than acquired antimicrobial resistance or elevated clinical risk (**33**).

Genomic hazard framework and implications

Integration of virulence, resistance, and plasmid mobility determinants into a composite framework indicated that most isolates recovered from the recycled water system were classified as low hazard, with a small proportion classified as moderate hazard. Moderate-hazard isolates were more frequent in post-treatment samples (W3) than in influent samples (W1), and the hazard-profile distribution differed significantly between these sites (Fisher's exact test, $P = 0.0012$). Although low-hazard isolates remained predominant at all sampling locations, the higher proportion of moderate-hazard isolates at W3 indicates that treatment-stage differences in the composition of the recovered *E. coli* population may warrant further consideration.

The predominance of low-hazard isolates is consistent with the generally low pathogenic potential expected among environmental *E. coli* populations and supports the interpretation that the *E. coli* population recovered from the recycled water system was predominantly composed of isolates with relatively low genomic hazard potential. Nevertheless, the detection of moderate-hazard isolates following treatment highlights the value of genomic surveillance approaches capable of identifying isolates with elevated virulence, resistance, or plasmid mobility characteristics that may not

be distinguished by conventional indicator enumeration alone. These findings demonstrate that genomic approaches can provide additional resolution beyond traditional culture-based monitoring by identifying less abundant isolates with potentially greater public health relevance. Integration of such frameworks into monitoring programs may therefore improve interpretation of *E. coli* detections and support more targeted hazard assessment and management (34, 35).

Limitations and future directions

Several limitations should be considered. First, this study was conducted at a single facility, and observed patterns may reflect site-specific conditions. Broader validation across systems and treatment configurations is needed to assess generalizability. Second, although W3 represents the most downstream stage of the treatment train and the final recycled water product, samples were collected from an effluent tap downstream of a closed contact basin following completion of all treatment processes. Consequently, the presence of isolates exhibiting genomic similarity to companion animal-, livestock-, or wildlife-associated contextual genomes cannot be attributed solely to treatment performance and may reflect a combination of persistence through treatment, prolonged hydraulic residence within the contact basin, or survival within downstream infrastructure such as biofilms. Distinguishing among these mechanisms was beyond the scope of the present study and would require targeted environmental sampling together with higher-resolution source-tracking approaches.

Third, the number of *E. coli* isolates recovered differed substantially among sampling locations, reflecting differences in *E. coli* abundance across treatment stages. In particular, only a limited number of isolates were recovered from W2 following advanced treatment and disinfection, restricting statistical comparisons involving this site. Although appropriate non-parametric statistical approaches were used and W2 was excluded from formal comparative analyses, the unequal sampling structure may have reduced statistical power and limited the representativeness of some comparisons. Consequently, observed differences in genomic diversity, source attribution, and hazard profiles should be interpreted as indicative trends within the sampled isolate collection rather than definitive population-level estimates.

Fourth, the isolate-based approach does not capture the full microbial community, including uncultivable organisms or low-abundance populations. Future studies incorporating shotgun metagenomics would provide a more comprehensive assessment of microbial diversity, antimicrobial resistance determinants, virulence-associated genes, and potential pathogen reservoirs within recycled water systems. Finally, the proposed framework represents a prototype approach and has not yet been validated against epidemiological outcomes, environmental surveillance data, or quantitative microbial risk assessment (QMRA) models. Future studies incorporating larger isolate collections, longitudinal sampling, phenotypic validation, and integration with exposure modeling will be important for refining and validating genomic hazard assessment approaches for water reuse applications.

Despite these limitations, the study highlights the potential value of integrating whole-genome sequencing into routine recycled water surveillance programs and provides a foundation for the development of a validated hazard-informed hazard assessment framework for water reuse systems.

Overall, genomic analysis of *E. coli* in a recycled water system revealed high lineage diversity, limited but detectable site-level structuring, and predominantly low-hazard populations, with a small proportion of moderate-hazard isolates detected across treatment stages. These findings indicate that genome-resolved analysis improves interpretation of *E. coli* detections beyond conventional enumeration and provides a practical basis for incorporating genomic data into hazard-informed monitoring of recycled water systems.

MATERIALS AND METHODS

Sample and strain collection

The study was conducted at a recycled water plant in Victoria, Australia, a 10,500-ha lagoon-based wastewater treatment and recycling facility that processes approximately 600 ML per day of predominantly domestic municipal wastewater, with additional urban and industrial trade waste inputs. The facility produces approximately 30,000 ML of recycled water annually for environmental and agricultural reuse. Although the influent does not receive direct agricultural runoff or concentrated livestock effluent streams, the site is situated within a mixed agricultural and wildlife landscape, with extensive livestock grazing areas and internationally significant wetlands supporting large populations of migratory and resident waterbirds. Consequently, opportunities for secondary environmental fecal deposition within open lagoon systems cannot be excluded.

Water samples (n = 154) were collected as grab samples between January 2024 and June 2025 from three locations within a recycled water treatment system: (i) treated wastewater influent entering the recycled water treatment system (W1, n = 36), (ii) post-treatment water following UV disinfection (W2a, n = 16) or UV disinfection followed by chlorination (W2b, n = 16), and (iii) final recycled water stored within the contact basin prior to distribution for Class A recycled water reuse applications (W3, n = 84) (Fig. 5; Table S8). In addition, two sludge-associated samples were collected from the W3 contact basin during routine maintenance activities (W3 sludge, n = 2).

Samples were processed following standard filtration procedures outlined in AS/NZS 4276.5:2019 at the One Water Laboratory (Monash University, Victoria, Australia). For *E. coli* isolation, 100 mL aliquots were filtered in duplicate through 0.45 µm membrane filters and cultured on MI agar (Edwards Group, Australia) at 37°C for 18–24 h. In parallel, positive Colilert trays obtained from routine *E. coli* quantification at W1, W2, and W3 by Australian Laboratory Services Global (Scoresby, VIC, Australia) were used as an additional source of *E. coli* isolates. Up to three presumptive *E. coli* colonies per sample were randomly selected from MI agar, subcultured onto MacConkey and XLD agar (Oxoid; BD Difco), and purified through successive streaking prior to downstream genomic analysis. All isolates were archived in the One Water Biobank under ISO9001-certified procedures.

DNA extraction and sequencing

Single colonies were cultured in LB broth and incubated at 37°C for 18–24 h, and genomic DNA was extracted using the Quick-DNA Miniprep Kit (Zymo Research, USA) according to the manufacturer's instructions. DNA was quantified using a Qubit fluorometer (Invitrogen), and ≥200 ng was used for library preparation.

Sequencing libraries were prepared using the Oxford Nanopore Rapid Barcoding Kit (SQK-RBK114, Oxford Nanopore, UK). Pooled libraries were purified and loaded onto an R10.4.1 MinION flow cell (FLO-MIN114) for 48 h at the Monash Genomics and Bioinformatics Platform. Base-calling of FAST5 files was performed using Dorado v0.5.1 in super-accuracy (SUP) mode.

Genome assembly and quality assessment

Raw read statistics were summarized using Seqkit v2.10.0 (36). *De novo* assembly for long-reads was performed with Dragonflye v1.2.1 (https://github.com/rpetit3/dragonflye) using default parameters, which included adaptor trimming, removal of reads <1,000 bp, assembly with Flye, polishing with Racon, and filtering short, low-coverage, or homopolymer-rich contigs. Final polishing was completed with Medaka v2.0.1 (https://github.com/nanoporetech/medaka) using the r1041_e82_400bps_sup_v5.0.0 model.

A minimum sequencing depth of >20× per genome was targeted. Recent benchmarking of bacterial SNP calling using Oxford Nanopore R10.4.1 data with super-accuracy (SUP) base calling has shown that accurate SNP inference can be achieved at ~10× coverage, with performance comparable to Illumina-based approaches (37). Accordingly,

sequencing depths in this study exceeded conservative thresholds reported to be sufficient for robust SNP-based analyses.

Assembly quality and completeness were assessed using CheckM v1.2.2 (≥90% completeness, <5% contamination) (38) and BUSCO v5.1.3 (Enterobacterales data set, ≥90% completeness) (39). Assembly statistics, including genome size, number of contigs, sequencing depth, GC content, and N50, were calculated using Seqkit v2.10.0 and are provided in Table S2. Taxonomic classification was performed with Kraken2 v2.1.3 (40) against a GTDB-derived database (downloaded May 2024). All analyses were conducted on the M3 High Performing Computing facility (Monash University, VIC, Australia) (https://research.monash.edu/en/equipments/high-performance-computing-m3massive).

Population genomics and phylogenetic analysis

In silico phylogroup assignment was performed using the Clermont Typing (http://clermontyping.iame-research.center/), and sequence types (STs) were determined by multi-locus sequence typing (MLST) v2.15 (Achtman scheme; https://github.com/tseemann/mlst). A contextual data set of 936 Australian *E. coli* genomes (clinical human n = 168, livestock_animal n = 239, wild_animal n = 107, companion animal n = 382, wastewater n = 40; Table S9) was retrieved from NCBI. Genomes were filtered using CheckM v1.2.2 (≥90% completeness, <5% contamination) (38), and taxonomic identity was confirmed using Kraken2 v2.1.3 (40). Pairwise genomic similarity was assessed using Mash v1.4.6 using default settings (41) and FastANI v1.33 (42). Of the 936 contextual genomes, 121 shared ≥99% ANI with at least one study isolate and were retained for comparative phylogenetic analysis.

Core genome alignments were generated with Roary v3.13.0 (43) using Prokka v1.14.6 annotation (.gff files) (44). Maximum-likelihood phylogenies were inferred using IQ-TREE v2.1.3 with 1,000 ultrafast bootstrap replicates (45). Phylogenetic trees were midpoint rooted for visualization purposes and displayed using iTOL v6 (46) alongside metadata (source attribution, collection date, ST, and phylogroup). For Australian reference genomes, source categories (human, companion animal, livestock, wild animal, and wastewater) were included as metadata tracks to facilitate interpretation of source attribution patterns and relationships with study isolates. SNP distances and clonal relationships were evaluated using SNP-sites v2.5.1 (47) to extract core SNP alignments and SNP-dists v0.6.3 (https://github.com/tseemann/snp-dists) to generate pairwise SNP distance matrices.

Development of genomic hazard framework

A genomic-based hazard assessment framework was developed to estimate the potential public health relevance of environmental *E. coli* isolates. The framework integrates four genomic dimensions: (i) virulence determinants, representing pathogenic potential; (ii) ARGs, representing the clinical significance of resistance; (iii) multidrug resistance (MDR) breadth, representing resistance diversity across antimicrobial classes; and (iv) plasmid-associated hazard, representing the potential for horizontal dissemination of resistance and virulence determinants. These components were analyzed as separate axes to preserve biological interpretability and were not initially collapsed into a single primary "hazard score." The framework is intended as a relative genomic screening tool for comparing isolates across treatment stages and exposure contexts.

Virulence factor identification and scoring

Virulence-associated genes were identified using ABRicate v1.0.1 (https://github.com/tseemann/abricate), against the VFDB database (48) (≥90% identity and ≥80% coverage). Genes were classified according to established *E. coli* pathotypes, including diarrheagenic *E. coli* (DEC) and extraintestinal pathogenic *E. coli* (ExPEC), including uropathogenic *E. coli* (UPEC).

Key DEC markers included genes associated with enteropathogenic/EPEC (*eae*, *bfpA),* Shiga toxin-producing/STEC and hemolytic uremic syndrome/HUS (*stx1*, *stx2*, *ehxA),* enteroaggregative/EAEC (*aggR*, *aaiC*, *aatA)*, enterotoxigenic/ETEC (*eltA/B*, *estA/B)*, enteroinvasive/EIEC, and diffusely adherent/DAEC (*ipaH*, *virF*, *afa*, *daaE*) (28, 29, 49).

Key ExPEC and uropathogenic *E. coli*/UPEC-associated markers included genes linked to adhesion, iron acquisition, and serum resistance (e.g., *papA/C*, *sfa/focDE*, *afa/draBC*, *iutA*, *chuA*, *kpsMII*, *fyuA*, *yfcV*, *vat)* (21, 50, 51), along with additional ExPEC-associated genes (*cnf*, *clbB*, *tcpC*, *usp*, *iss*, *traT*, *neuC*, *ibeA*, *iroN*, *iucC*, *irp2*, *sitA*) to support broader pathogenicity profiling (21, 29, 51, 52).

Virulence potential was quantified using a tiered scoring system (Table S10A). Isolates carrying major DEC markers or key ExPEC/UPEC-associated virulence determinants were assigned a high virulence score (3). Isolates with secondary or supporting virulence determinants were assigned an intermediate score (2), while those with only minor virulence genes were assigned a low score (1). Isolates lacking detectable virulence determinants were assigned a score of 0. The virulence score was defined as the highest category detected per isolate, representing maximal predicted pathogenic potential. Importantly, this framework reflects genomic pathogenic potential rather than confirmed clinical virulence.

Antimicrobial resistance gene detection and scoring

ARGs were detected using abritamr v1.0.14 (https://github.com/MDU-PHL/abritamr) with a threshold of ≥90% sequence identity and ≥80% gene coverage. Detected genes were classified according to antimicrobial class and clinical relevance, informed by the WHO Bacterial Priority Pathogens List (53). To focus the genomic hazard assessment on clinically relevant acquired antimicrobial resistance determinants, intrinsic chromosomal resistance-associated genes commonly present in *E. coli* were excluded from ARG hazard scoring. These included genes associated with multidrug efflux systems and their regulation (*acrAB-tolC*, *emrA/B*, *mdtM*, *marA*, *soxS*, *rob,* and *evgA*), outer cell membrane permeability and cell envelope structure (*ompC*, *ompF*, *rfaG,* and *lpxM*), and intrinsic *ampC* β-lactamase of *E. coli* (*bla*$_{EC}$) (54–57). These genes were therefore not considered acquired ARGs for hazard scoring purposes.

A hierarchical scoring system was applied to quantify ARG-associated genomic hazard based on clinical importance, mobility, and therapeutic relevance. High-hazard (score = 3) resistance determinants included carbapenemase genes (*bla*$_{NDM}$, *bla*$_{KPC}$, *bla*$_{OXA\text{-}48\text{-}like\ variants}$), ESBLs (e.g., *bla*$_{CTX\text{-}M}$ variants, *bla*$_{TEM\text{-}3/\text{-}10}$, *bla*$_{SHV}$ variants), and colistin resistance genes (*mcr-1* to *mcr-10*) (53, 58–60). These genes represent WHO critical-priority resistance determinants associated with multidrug-resistant infections and global dissemination (53, 61–63).

Moderate-hazard (score = 2) determinants included aminoglycoside resistance genes encoding 16S rRNA methyltransferases (e.g., *armA*, *rmtB*), conferring high-level aminoglycoside resistance (53, 64, 65), and fluoroquinolone resistance mediated by chromosomal quinolone resistance-determining region (QRDR) mutations. Isolates were assigned a score of 2 when they harbored ≥2 established QRDR mutations across *gyrA* and *parC*, consistent with clinically significant fluoroquinolone resistance in *E. coli*. Plasmid-mediated quinolone resistance (PMQR) genes, including *qnr* variants, *oqxAB, qepA*, and *aac*(6′)-Ib-cr, were not considered equivalent to QRDR-mediated resistance and were therefore classified separately as low-impact resistance markers (score = 1) when present. Additionally, low-hazard determinants (score = 1) included all remaining acquired ARGs not classified above. Isolates lacking detectable acquired ARGs were assigned a score of 0 (Table S10B).

To complement the ARG hazard score, a multidrug resistance (MDR) index was calculated to capture the breadth of antimicrobial resistance within each isolate. The MDR index was defined as the number of distinct antimicrobial classes for which at least one acquired resistance gene or relevant chromosomal mutation was detected per isolate. Each antimicrobial class was counted once per genome to avoid inflation due

to multiple genes within the same class. The MDR index therefore represents resistance diversity rather than gene abundance.

Plasmid-associated hazard assessment

Plasmid-associated contigs and mobility classes (conjugative, mobilizable, or non-mobilizable) were predicted using MOB-suite v3.1.9 under default parameters (66). To capture both horizontal gene transfer potential and gene carried by the plasmid, a composite plasmid hazard score was developed. Plasmid mobility (M) was used as a proxy for horizontal gene transfer potential. Conjugative plasmids were assigned a score of 2, mobilizable plasmids a score of 1, and non-mobilizable or absent plasmids a score of 0. Gene content (G) was independently scored according to the presence and clinical relevance of ARGs and virulence factors detected on plasmid-associated contigs. Plasmids carrying no detectable ARGs or virulence factors were assigned a score of 0; plasmids carrying virulence factors only or non-clinically relevant ARGs were assigned a score of 1; plasmids carrying clinically relevant ARGs were assigned a score of 2; and plasmids carrying both clinically relevant ARGs and virulence factors were assigned a score of 3 (Table S10C). The final plasmid hazard score (*P*) was calculated as the product of mobility and gene content scores ($P = M \times G$), thereby integrating both transfer potential and gene hazard into a single metric. Consequently, plasmids lacking detectable ARGs or virulence determinants received a plasmid hazard score of 0, irrespective of mobility classification, preventing highly mobile plasmids without relevant resistance or virulence cargo from being assigned elevated hazard scores. The final plasmid hazard score therefore ranged from 0 to 6, with higher scores indicating greater potential for dissemination of clinically relevant resistance and virulence determinants.

Standardized composite genomic hazard index (exploratory metric)

To enable integrated visualization and cross-isolate comparison, a standardized composite genomic hazard index (CGHI) was derived as an exploratory metric. This index was not used as a primary endpoint but served as a relative ranking and visualization tool. Each component was first normalized to a 0–1 scale:

$$\text{Virulence score:} \quad V_{\text{norm}} = \frac{V_{\text{score}}}{V_{\text{max}}}$$

$$\text{ARG hazard:} \quad A_{\text{norm}} = \frac{A_{\text{score}}}{A_{\text{max}}}$$

$$\text{MDR index:} \quad M_{\text{norm}} = \frac{M_{\text{score}}}{M_{\text{max}}}$$

$$\text{Plasmid hazard:} \quad P_{\text{norm}} = \frac{P_{\text{score}}}{P_{\text{max}}}$$

where V_{max}, A_{max}, and P_{max} represent the theoretical maximum possible scores for virulence (3), ARG (3), and plasmid-associated features (6), respectively. In contrast, the MDR index represents a count of resistance classes and does not have a predefined maximum within the scoring framework; M_{max} was therefore defined as the maximum observed MDR value within the data set ($M_{\text{max}} = 8$) and used for normalization.

The composite index was then calculated as the sum of the four normalized components using equal weighting:

$$\text{CGHI} = V_{\text{norm}} + A_{\text{norm}} + M_{\text{norm}} + P_{\text{norm}}$$

The CGHI ranged from 0 to 4, with higher values indicating a greater cumulative genomic hazard profile. Equal weighting was adopted as a pragmatic and transparent

approach because insufficient empirical evidence exists to support differential weighting of the four genomic dimensions. Importantly, the four components were also retained as independent analytical variables to preserve interpretability. The CGHI provides a supplementary integrative representation of genomic hazard for visualization, exploratory analysis, and relative ranking rather than a definitive measure of public health risk. The framework does not account for gene expression, phenotypic resistance, infectivity, host susceptibility, exposure frequency, infective dose, environmental persistence, or disease outcome and therefore should not be interpreted as a direct measure of human health risk. Accordingly, the CGHI is best interpreted as a measure of relative genomic hazard rather than quantitative human health risk.

Integration of genomic hazard framework with human exposure routes

To account for differences in potential human exposure, each water-use category was assigned an exposure score based on intended use and likely exposure pathways (Table S10D), following national and international guidelines for drinking, recycled, and recreational water (14, 15, 53, 67–69). Higher exposure scores indicate greater opportunities for human contact via ingestion, inhalation, or dermal exposure (70, 71). Exposure scores were converted to exposure weighting factors ranging from 0.25 to 1.0 (Table S10D). The weighting factors were applied to the composite genomic hazard index (CGHI) to generate an exposure-adjusted hazard index (EAHI):

$$\mathrm{EAHI} = \mathrm{CGHI} \times E_{\mathrm{norm}}$$

where E_{norm} is the exposure weighting factor. The EAHI ranged from 0 to 4, with higher values indicating isolates that combine greater genomic hazard with greater potential for human exposure (Table S10E). The EAHI was used as a relative ranking metric to contextualize genomic hazard according to plausible human exposure pathways and was not intended as a quantitative measure of public health risk.

Statistical analysis

All statistical analyses were conducted in R v4.4.0 (72). Differences in categorical variables were assessed using Fisher's exact or chi-square tests (73–75). Continuous variables were analyzed using Wilcoxon rank-sum or Kruskal–Wallis tests. *E. coli* concentrations were expressed as most probable number (MPN) per 100 mL and log-transformed as $\log_{10}(\mathrm{MPN} + 1)$ to account for the skewed distributions. As data were non-normally distributed (Shapiro–Wilk test, $P < 0.001$), non-parametric tests were used. Data visualization was performed using ggplot2 (75).

ACKNOWLEDGMENTS

This work was supported by the Medical Research Future Fund (MRFF; MRFCRI000195) and a Melbourne Water Research grant (LV0554-60243).

The authors gratefully acknowledge Georgia Lynch, Ravindran Appathurai, and the OWL and ALS teams for their technical support and assistance with sample collection.

The authors used ChatGPT (OpenAI) to assist with code generation and data visualization and Grammarly to support language editing. All outputs were independently reviewed and validated by the authors, who take full responsibility for the accuracy, integrity, and originality of the work.

AUTHOR AFFILIATIONS

[1]School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Victoria, Australia
[2]Melbourne Water Corporation, Melbourne, Victoria, Australia

[3]Melbourne Veterinary School, Faculty of Science, University of Melbourne, Parkville, Victoria, Australia
[4]One Water Laboratory, Department of Civil and Environmental Engineering, Monash University, Clayton, Victoria, Australia

AUTHOR ORCIDs

Dewa A. P. Rasmika Dewi http://orcid.org/0000-0003-2775-9079

FUNDING

Funder	Grant(s)	Author(s)
Medical Research Future Fund (MRFF), Australian Government	MRFCRI000195	Dewa A. P. Rasmika Dewi
Medical Research Future Fund (MRFF), Australian Government	MRFCRI000195	Rebekah Henry
Melbourne Water	LV0554-60245	Dewa A. P. Rasmika Dewi Rebekah Henry

AUTHOR CONTRIBUTIONS

Dewa A. P. Rasmika Dewi, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review and editing | Harshanie Abeywardena, Data curation, Methodology, Project administration, Resources, Writing – review and editing | Dieter Bulach, Methodology, Software, Writing – review and editing | Chi-Wen Tseng, Investigation, Writing – review and editing | Lamiya Bata, Investigation, Writing – review and editing | Fiona Lynch, Supervision, Writing – review and editing | Melita Stevens, Supervision, Writing – review and editing | Nicholas D. Crosbie, Conceptualization, Methodology, Resources, Supervision, Writing – review and editing | Rebekah Henry, Conceptualization, Methodology, Supervision, Writing – review and editing

DATA AVAILABILITY

Raw sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1369737, with associated BioSample records SAMN53398571–SAMN53398754.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental material (AEM01342-26-s0001.docx). Tables S1, S5, and S10; Fig. S1 to S4.
Table S2 (AEM01342-26-s0002.xlsx). Genomic characteristics and population structure of *Escherichia coli* isolates recovered from the recycled water system.
Table S3 (AEM01342-26-s0003.xlsb). Gene content of *Escherichia coli* genomes recovered from the recycled water system.
Table S4 (AEM01342-26-s0004.xlsx). ANI-based contextual genome comparisons and source attribution of *Escherichia coli* isolates recovered from the recycled water system.
Table S6 (AEM01342-26-s0005.xlsx). Virulence factors, antimicrobial resistance determinants, and plasmid-associated features of *Escherichia coli* isolates recovered from the recycled water system.
Table S7 (AEM01342-26-s0006.xlsx). Genomic hazard scoring framework for *Escherichia coli* isolates recovered from the recycled water system.

Table S8 (AEM01342-26-s0007.xlsx). Sample and isolate metadata for *Escherichia coli* recovered from the recycled water system.
Table S9 (AEM01342-26-s0008.xlsx). Contextual reference genome metadata used for comparative genomic and source attribution analyses of *Escherichia coli* isolates.

REFERENCES

1. Shemer H, Wald S, Semiat R. 2023. Challenges and solutions for global water scarcity. Membranes (Basel) 13:612. https://doi.org/10.3390/membranes13060612
2. Radcliffe JC. 2022. Current status of recycled water for agricultural irrigation in Australia, potential opportunities and areas of emerging concern. Sci Total Environ 807:151676. https://doi.org/10.1016/j.scitotenv.2021.151676
3. Foster-Nyarko E, Pallen MJ. 2022. The microbial ecology of *Escherichia coli* in the vertebrate gut. FEMS Microbiol Rev 46:fuac008. https://doi.org/10.1093/femsre/fuac008
4. Nowicki S, deLaurent ZR, de Villiers EP, Githinji G, Charles KJ. 2021. The utility of *Escherichia coli* as a contamination indicator for rural drinking water: evidence from whole genome sequencing. PLoS One 16:e0245910. https://doi.org/10.1371/journal.pone.0245910
5. Kim HJ, Yoon HW, Lee MA, Kim YH, Lee CJ. 2023. Impact of UV-C irradiation on bacterial disinfection in a drinking water purification system. J Microbiol Biotechnol 33:106–113. https://doi.org/10.4014/jmb.2211.11027
6. Koh XP, Shen Z, Woo CF, Yu Y, Lun HI, Cheung SW, Kwan JKC, Lau SCK. 2022. Genetic and ecological diversity of *Escherichia coli* and cryptic *Escherichia* clades in subtropical aquatic environments. Front Microbiol 13:811755. https://doi.org/10.3389/fmicb.2022.811755
7. Rumball NA, Alm EW, McLellan SL. 2023. Genetic determinants of *Escherichia coli* survival in beach sand. Appl Environ Microbiol 89:e0142322. https://doi.org/10.1128/aem.01423-22
8. Vogeleer P, Tremblay YDN, Mafu AA, Jacques M, Harel J. 2014. Life on the outside: role of biofilms in environmental persistence of Shiga-toxin producing *Escherichia coli*. Front Microbiol 5:317. https://doi.org/10.3389/fmicb.2014.00317
9. Dang C, Xia Y, Zheng M, Liu T, Liu W, Chen Q, Ni J. 2020. Metagenomic insights into the profile of antibiotic resistomes in a large drinking water reservoir. Environ Int 136:105449. https://doi.org/10.1016/j.envint.2019.105449
10. Luo G, Li B, Li LG, Zhang T, Angelidaki I. 2017. Antibiotic resistance genes and correlations with microbial community and metal resistance genes in full-scale biogas reactors as revealed by metagenomic analysis. Environ Sci Technol 51:4069–4080. https://doi.org/10.1021/acs.est.6b05100
11. Tarek MH, Garner E. 2023. A proposed framework for the identification of indicator genes for monitoring antibiotic resistance in wastewater: insights from metagenomic sequencing. Sci Total Environ 854:158698. https://doi.org/10.1016/j.scitotenv.2022.158698
12. Xu X, Lin Y, Deng Y, Liu L, Wang D, Tang Q, Wang C, Chen X, Che Y, Wyrsch ER, Jarocki VM, Djordjevic SP, Zhang T. 2025. Ecological connectivity of genomic markers of antimicrobial resistance in *Escherichia coli* in Hong Kong. Nat Commun 16:7319. https://doi.org/10.1038/s41467-025-62455-w
13. Odonkor ST, Mahami T. 2020. *Escherichia coli* as a tool for disease risk assessment of drinking water sources. Int J Microbiol 2020:2534130. https://doi.org/10.1155/2020/2534130
14. Health-Victoria. 2024. Class a recycled water. Department of Health Victoria. Available from: https://www.health.vic.gov.au/water/class-a-recycled-water. Retrieved 8 Oct 2025.
15. Environment Protection Authority Victoria. 2021. Victorian guideline for water recycling. Melbourne, Australia Environment Protection Authority Victoria. https://www.epa.vic.gov.au/sites/default/files/2025-05/Victorian-guideline-for-water-recycling.pdf.
16. Reitter C, Petzoldt H, Korth A, Schwab F, Stange C, Hambsch B, Tiehm A, Lagkouvardos I, Gescher J, Hügler M. 2021. Seasonal dynamics in the number and composition of coliform bacteria in drinking water reservoirs. Sci Total Environ 787:147539. https://doi.org/10.1016/j.scitotenv.2021.147539
17. Abia ALK, Ubomba-Jaswa E, Momba MNB. 2015. Impact of seasonal variation on *Escherichia coli* concentrations in the riverbed sediments in the Apies River, South Africa. Sci Total Environ 537:462–469. https://doi.org/10.1016/j.scitotenv.2015.07.132
18. Aguirre-Sánchez JR, Valdez-Torres JB, Del Campo NC, Martínez-Urtaza J, Del Campo NC, Lee BG, Quiñones B, Chaidez-Quiroz C. 2022. Phylogenetic group and virulence profile classification in *Escherichia coli* from distinct isolation sources in Mexico. Infect Genet Evol 106:105380. https://doi.org/10.1016/j.meegid.2022.105380
19. Saraceno M, Gómez Lugo S, Ortiz N, Gómez BM, Sabio Y García CA, Frankel N, Graziano M. 2021. Unraveling the ecological processes modulating the population structure of *Escherichia coli* in a highly polluted urban stream network. Sci Rep 11:14679. https://doi.org/10.1038/s41598-021-94198-1
20. Stoppe N de C, Silva JS, Carlos C, Sato MIZ, Saraiva AM, Ottoboni LMM, Torres TT. 2017. Worldwide phylogenetic group patterns of *Escherichia coli* from commensal human and wastewater treatment plant isolates. Front Microbiol 8:2512. https://doi.org/10.3389/fmicb.2017.02512
21. Lagerstrom KM, Scales NC, Hadly EA. 2024. Impressive pan-genomic diversity of *E. coli* from a wild animal community near urban development reflects human impacts. iScience 27:109072. https://doi.org/10.1016/j.isci.2024.109072
22. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, Calland JK, Yahara K, Murray S, Wilkinson TS, Williams LK, Hitchings MD, Porter J, Kemmett K, Feil EJ, Jolley KA, Williams NJ, Corander J, Sheppard SK. 2021. Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. Nat Commun 12:765. https://doi.org/10.1038/s41467-021-20988-w
23. Ma Y, Qiao Y, Zhang X, Ye L. 2024. Filamentous bacteria-induced sludge bulking can alter antibiotic resistance gene profiles and increase potential risks in wastewater treatment systems. Environ Int 190:108920. https://doi.org/10.1016/j.envint.2024.108920
24. Zhao B, Zhang R, Jin B, Yu Z, Wen W, Zhao T, Quan Y, Zhou J. 2025. Sludge water: a potential pathway for the spread of antibiotic resistance and pathogenic bacteria from hospitals to the environment. Front Microbiol 16:1492128. https://doi.org/10.3389/fmicb.2025.1492128
25. Muloi DM, Wee BA, McClean DMH, Ward MJ, Pankhurst L, Phan H, Ivens AC, Kivali V, Kiyong'a A, Ndinda C, et al. 2022. Population genomics of *Escherichia coli* in livestock-keeping households across a rapidly developing urban landscape. Nat Microbiol 7:581–589. https://doi.org/10.1038/s41564-022-01079-y
26. White AP, Sibley KA, Sibley CD, Wasmuth JD, Schaefer R, Surette MG, Edge TA, Neumann NF. 2011. Intergenic sequence comparison of *Escherichia coli* isolates reveals lifestyle adaptations but not host specificity. Appl Environ Microbiol 77:7620–7632. https://doi.org/10.1128/AEM.05909-11
27. Zhou Z, Charlesworth J, Achtman M. 2020. The EnteroBase user guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. Genome Res 30:138–152. https://doi.org/10.1101/gr.251678.119
28. Pakbin B, Brück WM, Rossen JWA. 2021. Virulence factors of enteric pathogenic *Escherichia coli*: a review. IJMS 22:9922. https://doi.org/10.3390/ijms22189922
29. Vanstokstraeten R, Moral C, Wybo I, Geyter D, Janssen T, Caljon B, Demuyser T, Crombe F, Pierard D. 2022. Molecular characterization of extraintestinal and diarrheagenic *Escherichia coli* blood isolates. Virulence 13:2032–2041. https://doi.org/10.1080/21505594.2022.2147735
30. Abberton CL, Bereschenko L, van der Wielen PWJJ, Smith CJ. 2016. Survival, biofilm formation, and growth potential of environmental and enteric *Escherichia coli* strains in drinking water microcosms. Appl Environ Microbiol 82:5320–5331. https://doi.org/10.1128/AEM.01569-16
31. Kang D, Kirienko NV. 2018. Interdependence between iron acquisition and biofilm formation in *Pseudomonas aeruginosa*. J Microbiol 56:449–457. https://doi.org/10.1007/s12275-018-8114-3

32. Zhou F, Wang D, Hu J, Zhang Y, Tan BK, Lin S. 2022. Control measurements of *Escherichia coli* biofilm: a review. Foods 11:2469. https://doi.org/10.3390/foods11162469
33. Du D, Wang-Kan X, Neuberger A, van Veen HW, Pos KM, Piddock LJV, Luisi BF. 2018. Multidrug efflux pumps: structure, function and regulation. Nat Rev Microbiol 16:523–539. https://doi.org/10.1038/s41579-018-0048-6
34. de Brito Cruz D, Schmidt PJ, Emelko MB. 2024. Drinking water QMRA and decision-making: sensitivity of risk to common independence assumptions about model inputs. Water Res 259:121877. https://doi.org/10.1016/j.watres.2024.121877
35. Morales-Mora E, Rivera-Montero L, Montiel-Mora JR, Barrantes-Jiménez K, Chacón-Jiménez L. 2025. Assessing microbial risks of *Escherichia coli*: a spatial and temporal study of virulence and resistance genes in surface water in resource-limited regions. Sci Total Environ 958:178044. https://doi.org/10.1016/j.scitotenv.2024.178044
36. Shen W, Sipos B, Zhao L. 2024. SeqKit2: a Swiss army knife for sequence and alignment processing. iMeta 3:e191. https://doi.org/10.1002/imt2.191
37. Hall MB, Wick RR, Judd LM, Nguyen AN, Steinig EJ, Xie O, Davies M, Seemann T, Stinear TP, Coin L. 2024. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data. eLife 13:RP98300. https://doi.org/10.7554/eLife.98300
38. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055. https://doi.org/10.1101/gr.186072.114
39. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2021. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 39:105–114. https://doi.org/10.1038/s41587-020-0603-3
40. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0
41. Katz LS, Griswold T, Morrison SS, Caravas JA, Zhang S, den Bakker HC, Deng X, Carleton HA. 2019. Mashtree: a rapid comparison of whole genome sequence files. J Open Source Softw 4:44. https://doi.org/10.21105/joss.01762
42. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:5114. https://doi.org/10.1038/s41467-018-07641-9
43. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421
44. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153
45. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
46. Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res 52:W78–W82. https://doi.org/10.1093/nar/gkae268
47. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. *SNP-sites*: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mgen.0.000056
48. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res 44:D694–D697. https://doi.org/10.1093/nar/gkv1239
49. Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic *Escherichia coli*. Nat Rev Microbiol 2:123–140. https://doi.org/10.1038/nrmicro818
50. Malberg Tetzschner AM, Johnson JR, Johnston BD, Lund O, Scheutz F. 2020. *In silico* genotyping of *Escherichia coli* isolates for extraintestinal virulence genes by use of whole-genome sequencing data. J Clin Microbiol 58:10. https://doi.org/10.1128/JCM.01269-20
51. Sora VM, Meroni G, Martino PA, Soggiu A, Bonizzi L, Zecconi A. 2021. Extraintestinal pathogenic *Escherichia coli*: virulence factors and antibiotic resistance. Pathogens 10:1355. https://doi.org/10.3390/pathogens10111355
52. Johnson JR, Johnston BD, Porter S, Thuras P, Aziz M, Price LB. 2019. Accessory traits and phylogenetic background predict *Escherichia coli* extraintestinal virulence better than does ecological source. J Infect Dis 219:121–132. https://doi.org/10.1093/infdis/jiy459
53. World Health Organization. 2024. WHO bacterial priority pathogens list, 2024: Bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance. Geneva World Health Organization. https://www.who.int/publications/i/item/9789240093461.
54. Poirel L, Madec J, Lupo A, Schink A, Kieffer N, Nordmann P. 2018. Antimicrobial resistance in *Escherichia coli*. Microbiol Spectr 6. https://doi.org/10.1128/microbiolspec.ARBA-0026-2017
55. Nasrollahian S, Graham JP, Halaji M. 2024. A review of the mechanisms that confer antibiotic resistance in pathotypes of *E. coli*. Front Cell Infect Microbiol 14:1387497. https://doi.org/10.3389/fcimb.2024.1387497
56. Balachandran M, Chatterjee R, Chaudhary I, Jena C, Matange N. 2025. Impeding pathways of intrinsic resistance in *Escherichia coli* confers antibiotic sensitization and resistance proofing. PLoS Biol 23:e3003443. https://doi.org/10.1371/journal.pbio.3003443
57. Thakur D, Kumar L. 2026. Biofilm-associated *Escherichia coli* infections: pathogenesis, clinical implications, and treatment strategies. Crit Rev Microbiol 52:198–242. https://doi.org/10.1080/1040841X.2025.2548490
58. Aljohni MS, Harun-Ur-Rashid M, Selim S. 2025. Emerging threats: antimicrobial resistance in extended-spectrum beta-lactamase and carbapenem-resistant *Escherichia coli*. Microb Pathog 200:107275. https://doi.org/10.1016/j.micpath.2024.107275
59. Sivarajan V, Ganesh AV, Subramani P, Ganesapandi P, Sivanandan RN, Prakash S, Manikandan N, Dharmarajan A, Arfuso F, Warrier S, Raj M, Perumal K. 2025. Prevalence and genomic insights of carbapenem resistant and ESBL producing Multidrug resistant *Escherichia coli* in urinary tract infections. Sci Rep 15:2541. https://doi.org/10.1038/s41598-024-84754-w
60. Liu JH, Liu YY, Shen YB, Yang J, Walsh TR, Wang Y, Shen J. 2024. Plasmid-mediated colistin-resistance genes: mcr. Trends Microbiol 32:365–378. https://doi.org/10.1016/j.tim.2023.10.006
61. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X, Yu L-F, Gu D, Ren H, Chen X, Lv L, He D, Zhou H, Liang Z, Liu J-H, Shen J. 2016. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis 16:161–168. https://doi.org/10.1016/S1473-3099(15)00424-7
62. Logan LK, Weinstein RA. 2017. The epidemiology of carbapenem-resistant enterobacteriaceae: the impact and evolution of a global menace. J Infect Dis 215:S28–S36. https://doi.org/10.1093/infdis/jiw282
63. Nordmann P, Naas T, Poirel L. 2011. Global spread of carbapenemase-producing *Enterobacteriaceae*. Emerg Infect Dis 17:1791–1798. https://doi.org/10.3201/eid1710.110655
64. Doma AO, Popescu R, Mituleţu M, Muntean D, Dégi J, Boldea MV, Radulov I, Dumitrescu E, Muselin F, Puvača N, Cristina RT. 2020. Comparative evaluation of qnrA, qnrB, and qnrS genes in *Enterobacteriaceae* ciprofloxacin-resistant cases, in swine units and a hospital from Western Romania. Antibiotics 9:698. https://doi.org/10.3390/antibiotics9100698
65. Kherroubi L, Bacon J, Rahman KM. 2024. Navigating fluoroquinolone resistance in Gram-negative bacteria: a comprehensive evaluation. JAC-Antimicrobial Resistance 6:dlae127. https://doi.org/10.1093/jacamr/dlae127
66. Robertson J, Nash JHE. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom 4. https://doi.org/10.1099/mgen.0.000206
67. National Health and Medical Research Council. 2008. Australian guidelines for water recycling: managing health and environmental risks (phase2). Canberra, Australia National Health and Medical Research Council. https://wwWnhmrc.gov.au/about-us/publications/australian-guidelines-water-recycling.
68. National Health and Medical Research Council. 2011. National water quality management strategy: australian drinking water guidelines 6. Canberra, Australia National Health and Medical Research Council. https://www.nhmrc.gov.au/about-us/publications/australian-drinking-water-guidelines.
69. World Health Organization. 2024. Guidelines for drinking-water quality: small water supplies. World Health Organization, Switzerland. https://www.who.int/publications/i/item/9789240088740.

70. Gerdes ME, Miko S, Kunz JM, Hannapel EJ, Hlavsa MC, Hughes MJ, Stuckey MJ, Francois Watkins LK, Cope JR, Yoder JS, Hill VR, Collier SA. 2023. Estimating waterborne infectious disease burden by exposure route, United States, 2014. Emerg Infect Dis 29. https://doi.org/10.3201/eid2907.230231
71. O'Flaherty E, Solimini A, Pantanella F, Cummins E. 2019. The potential human exposure to antibiotic resistant-*Escherichia coli* through recreational water. Sci Total Environ 650:786–795. https://doi.org/10.1016/j.scitotenv.2018.09.018
72. R-Core-TeaM. 2021. R: a language and environment for statistical computing. R Foundation for Statistical Computing
73. Agresti A. 2007. An introduction to categorical data analysis. John Wiley & Sons.
74. Fay MP. 2010. Confidence intervals that match Fisher's exact or Blaker's exact tests. Biostatistics 11:373–374. https://doi.org/10.1093/biostatistics/kxp050
75. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag.