Research Square

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

# Whole-genome sequence data of an ESBL-producing Escherichia coli from tap water in South Dhaka, Bangladesh

**Shibbir Ahmed**
Jagannath University
**Rasel Khan Nishan**
Jagannath University
**Jannatul Ferdous**
Jagannath University
**Abha Rahman**
Jagannath University
**Jannatul Ferdous.**
Independent University, Bangladesh
**Ashrafus Safa**
Independent University, Bangladesh
**Arafat Al Mamun**
University of Dhaka
**Md. Shariful Islam**
s_islam@mib.jnu.ac.bd

Jagannath University
**Abul Kalam Azad**
akazad88@mib.jnu.ac.bd

Jagannath University

**Data Note**

**Keywords:**

**Posted Date:** November 12th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11694215/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Whole-genome sequence data of an ESBL-producing Escherichia coli from tap water in South Dhaka, Bangladesh

Shibbir Ahmed[1], Rasel Khan Nishan[1], Jannatul Ferdous[1], Abha Rahman[1], Jannatul Ferdous[2], Ashrafus Safa[2], Arafat Al Mamun[3], Md. Shariful Islam[1*], Abul Kalam Azad[1*]

[1]Department of Microbiology, Jagannath University, Dhaka-1100, Bangladesh.

[2]Department of Life Sciences, School of Environment and Life Sciences, Independent University, Bangladesh, Dhaka-1245, Bangladesh.

[3]Centre for Advanced Research in Sciences (CARS), University of Dhaka, Dhaka-1000, Bangladesh.

*Corresponding authors: [Abul Kalam Azad, email: akazad88@mib.jnu.ac.bd], [Md. Shariful Islam, email: s_islam@mib.jnu.ac.bd]

Abstract

Escherichia coli (E. coli) harboring extended-spectrum beta-lactamase (ESBL) enzymes poses a significant antimicrobial resistance threat, compromising β-lactam efficacy. Here, we report the draft genome and genomic characterization of ESBL-producing E. coli strain S11, isolated from municipal tap water in South Dhaka, Bangladesh, sequenced using the Illumina NextSeq 2000 platform. The draft assembly comprised 4,990,311 bp distributed across 139 contigs of at least 200 bp, with an N50 value of 182,030 bp, an N90 value of 54,615 bp, a largest contig of 543,452 bp, and a GC content of 50.52% and an estimated nominal sequencing coverage of ~202.9×. ABRicate/CARD analysis identified 48 antimicrobial resistance-associated determinants, including intrinsic efflux and regulatory loci. Resistome screening identified the ESBL-associated gene blaCTX-M-15 together with other beta-lactamase and aminoglycoside resistance genes, including blaTEM-1B, blaOXA-1, blaCMY-145, aac(3)-IIa, and aac(6')-Ib-cr. PlasmidFinder detected Col156, Col8282, and IncI1 replicons. PathogenFinder2 predicted human pathogenic potential with a mean score of 0.9807. Collectively, these genomic findings indicate that this isolate carries antimicrobial resistance determinants and plasmid replicons, providing genomic data that may be useful for future comparative studies of environmental *E. coli*. Raw reads of strain S11 are available in NCBI under SRA SRR39239575, BioProject PRJNA1478760 and BioSample SAMN61018556 and assembly is available as GenBank WGS project JCAKBG000000000.

Specifications Table

| | |
|---|---|
| **Subject** | Bacterial sequencing |
| **Specific subject area** | Whole genome sequencing of an ESBL-producing *Escherichia coli* from a tap water source |
| **Type of data** | Raw sequencing reads, draft assembly, annotation files and derived analysis tables |
| **Data collection** | Illumina NextSeq 2000 platform (paired-end 2 × 151 bp; Illumina DNA Prep / Nextera DNA Flex library kit) |
| **Data source location** | City/Town/Region: South Dhaka, Dhaka Division<br>**Country:** Bangladesh<br>Latitude/longitude: 23°43'21.9"N 90°24'37.9"E |
| **Data accessibility** | Data is hosted on NCBI<br>**BioProject:** PRJNA1478760 **(https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1478760)**<br>**BioSample:** SAMN61018556 (**https://www.ncbi.nlm.nih.gov/biosample/SAMN61018556)**<br>**SRA:** SRR39239575 **(https://www.ncbi.nlm.nih.gov/sra/SRR39239575)**<br>**GenBank WGS project**: JCAKBG000000000 (**https://www.ncbi.nlm.nih.gov/nuccore/JCAKBG000000000**) |
| **Related research article** | None |

## Value of the Data

- The draft genome sequence of ESBL-producing *Escherichia coli* strain S11 provides genomic information on an antimicrobial-resistant *E. coli* isolate recovered from tap water in South Dhaka, Bangladesh.
- The dataset contains six acquired resistance-gene hits in ResFinder; the broader CARD screen contains 48 resistance-associated gene-locus hits, including intrinsic and regulatory loci.
- The genome provides useful information for comparing waterborne *E. coli* isolates with clinical, environmental, foodborne, and animal-associated strains within a One Health antimicrobial resistance framework.
- These data (BioProject PRJNA1478760) provide a single environmental genome that may support comparative analyses of ESBL-producing E. coli from urban water sources.

## Background

*Escherichia coli* is a Gram-negative bacterium commonly associated with the intestinal tract of humans and animals. Although many *E. coli* strains are commensal, some lineages carry virulence factors and antimicrobial resistance determinants that enable them to cause intestinal or extraintestinal infections. The occurrence of ESBL-producing *E. coli* is a major public health concern because ESBL enzymes can hydrolyze extended-spectrum cephalosporins and aztreonam, limiting therapeutic options and facilitating the spread of beta-lactam resistance [1, 2].

Water is an important environmental interface connecting humans, animals, sewage, and the broader community. Detection of ESBL-producing E. coli in tap water may indicate contamination and highlights the importance of monitoring antimicrobial resistance determinants in water systems. Whole-genome sequencing enables high-resolution characterization of resistant bacteria by identifying antimicrobial resistance genes, virulence-associated genes, plasmid replicons, serotype, phylogroup, and genome quality features. Genomic characterization of ESBL-producing *E. coli* from tap water in Bangladesh can therefore provide useful data for antimicrobial resistance surveillance and One Health-based public health monitoring [3, 4, 5]. Detailed sample metadata for E. coli strain S11 are presented in Supplementary File 1.

## Data Description

This study reports the draft genome sequence and in silico genomic characterization of ESBL-producing Escherichia coli strain S11, isolated from tap water in South Dhaka, Bangladesh. ANI-based taxonomic analysis confirmed the isolate as E. coli, with a highest ANI of 98.4463% against E. coli SMS-3-5 (GCA_000019645.1), query coverage of 86.17% and reference coverage of 82.46%, exceeding the 96% ANI cutoff used for species confirmation [6]. The final draft assembly comprised 4,990,311 bp distributed across 139 contigs, with a GC content of 50.52% and an N50

value of 182,030 bp. Genome quality assessment indicated 100.0% completeness with 0.09% contamination, supporting the reliability of the assembly for downstream genomic analysis. Genome annotation identified 4,991 genes, including 4,701 protein coding genes, 108 RNA genes (87 tRNAs, 16 rRNAs and 5 ncRNAs) and 182 pseudogenes. The circular genome map generated using Proksee provides a visual overview of GC content, GC skew, coding regions, and CRISPR features across the draft genome (**Fig. 1**). In silico typing assigned strain S11 to serotype O9:H4 and Clermont phylogroup F. The O-antigen-associated genes wzm and wzt, together with the H-antigen gene fliC, showed complete identity and coverage in the ECTyper output. Assembly, sequencing and annotation characteristics of Escherichia coli strain S11 are summarized in Table 1.

Table 1. Assembly, sequencing and annotation characteristics of Escherichia coli strain S11.

| **Characteristic** | **Updated value** |
|---|---|
| Assembly length (bp) | 4,990,311 |
| Contigs ≥200 bp | 139 |
| GC content (%) | 50.52 |
| N50 (bp) | 182,030 |
| N90 (bp) | 54,615 |
| Largest contig (bp) | 543,452 |
| L50 / L90 | 9 / 30 |
| Nominal depth (×) | 202.9 |
| Shovill subsampling target (×) | 150 |
| CheckM2 completeness (%) | 100.0 |
| CheckM2 contamination (%) | 0.09 |
| PGAP genes (total) | 4,991 |
| Protein-coding genes | 4,701 |
| RNA genes | 108 (87 tRNAs, 16 rRNAs, 5 ncRNAs) |
| Pseudogenes | 182 |
| CDSs (total, PGAP summary) | 4,883 |

Genome size, contig number and GC content refer to all 139 contigs ≥200 bp. GC content was calculated directly from this FASTA (50.51565%, rounded to 50.52%); QUAST reports 50.51%

for its ≥500-bp subset. N50, N90 and largest contig are the same for both thresholds. Nominal depth was calculated as retained bases divided by full assembly length. PGAP gene categories sum to 4,991; total CDSs include pseudogene CDSs. Gene totals follow the PGAP structured annotation summary.

Resistome screening identified acquired resistance-gene matches and broader resistance-associated loci. ResFinder detected six acquired resistance genes, including the ESBL-associated gene blaCTX-M-15, along with blaTEM-1B, blaOXA-1, blaCMY-145, aac(3)-IIa, and aac(6'')-Ib-cr. The relationship between these acquired resistance genes and their corresponding antimicrobial classes is shown in (Fig. 2. The aac(6')-Ib-cr acetyltransferase is associated primarily with aminoglycoside modification and also with reduced fluoroquinolone susceptibility. AMRFinderPlus report contained 14 core AMR-associated entries comprising seven gene-associated matches and seven distinct resistance-associated amino-acid changes. The gene-associated matches are blaTEM-1, blaCTX-M-15, blaOXA-1, blaCMY-145, aac(3)-IIe, aac(6')-Ib-cr5 and a partial catB3 hit with 70% reference coverage. The quinolone-associated entries are gyrA_S83L, gyrA_D87N, parC_S80I and parE_S458A; the remaining changes are nfsA_G126R, ftsI_N337NYRIN and cyaA_S352T (Supplementary Table S1). ABRicate with the NCBI database reported seven hits. ABRicate/CARD detected 48 resistance-associated gene-locus hits. Their associated mechanisms were antibiotic efflux (37 hits), antibiotic inactivation (7), antibiotic target alteration (3) and reduced permeability to antibiotic (2) (Fig. 3). All AMR hits are provided in Supplementary Table S2, and the curated resistance-associated changes are listed in Supplementary Table S1. PlasmidFinder detected three plasmid replicon types, Col156, Col8282, and IncI1,. None of the six ResFinder hits was on a contig containing a replicon hit in either the native PlasmidFinder or ABRicate/plasmidfinder report. blaCTX-M-15 was located on contig00051 at positions 2729–3604, with an annotated Tn3-family transposase pseudogene at 96–2324 and a partial IS1380-family transposase at 3860–4055. Replicon hits and contig-context comparisons are provided in Supplementary Tables S3a and S3b. Virulence-associated screening using ABRicate detected 99 VFDB hits and 182 ecoli_vf hits, including loci related to adhesion, iron acquisition and stress response. The VFDB report contained 65 hits with full reference coverage and 34 with partial reference coverage; the ecoli_vf report contained 166 and 16, respectively. All 281 ABRicate database-specific virulence hits are provided in Supplementary Table S4. Standalone VirulenceFinder additionally reported 20 database matches corresponding to 19 gene-locus combinations and 17 distinct gene names (Supplementary Table S5). Seventeen matches covered the full reference and three were partial. Two eilA reference matches corresponded to the same genomic locus. PathogenFinder2 v0.6.0 predicted human pathogenic potential with a mean score of 0.9807 (Supplementary Table S6) [18].

Overall, the draft genome of E. coli strain S11 provides genomic data from a single ESBL-producing E. coli recovered from tap water. The recorded AST categories indicate resistance to multiple antimicrobial classes (Supplementary Table S7), while the genome contains blaCTX-M-15, additional resistance-gene matches, plasmid replicons and virulence-associated loci.

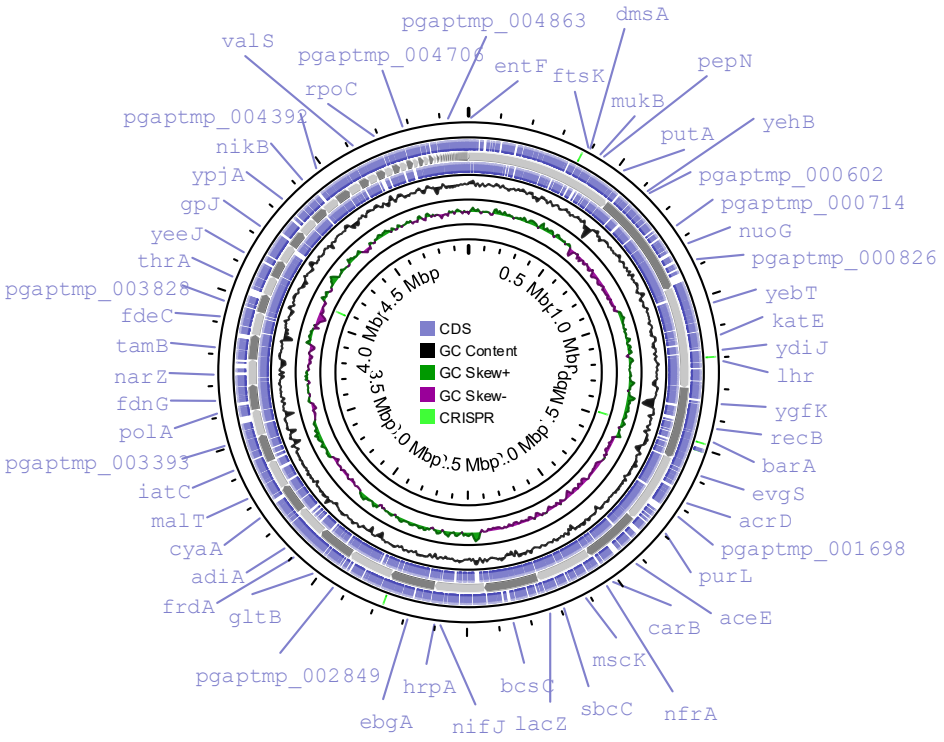



Figure 1. Circular genome map of *Escherichia coli* strain S11. The circular map shows the genomic features of the 139-contig draft assembly, including GC content, positive and negative GC skew, coding sequences, and CRISPR-associated features.

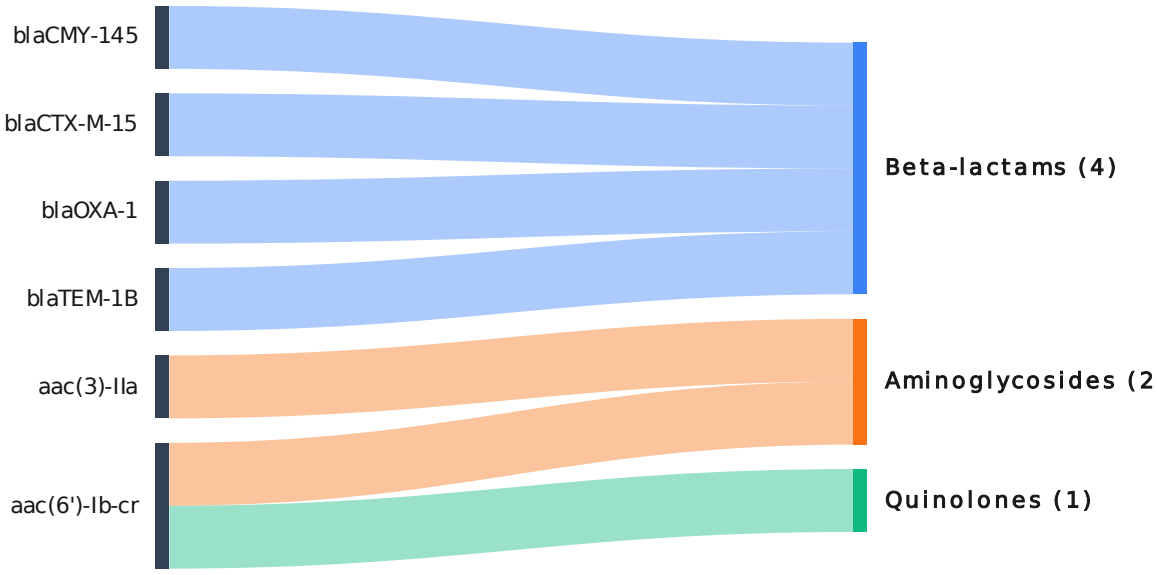


Figure 2. Acquired antimicrobial resistance gene profile of ESBL-producing Escherichia coli strain S11. Sankey diagram showing six genes and seven gene–class associations based on ResFinder/ABRicate analysis. Hits with coverage ≥90% and identity ≥80% were included. The aac(6')-Ib-cr gene contributes to both aminoglycoside and quinolone classes, giving four beta-lactam, two aminoglycoside and one quinolone associations.

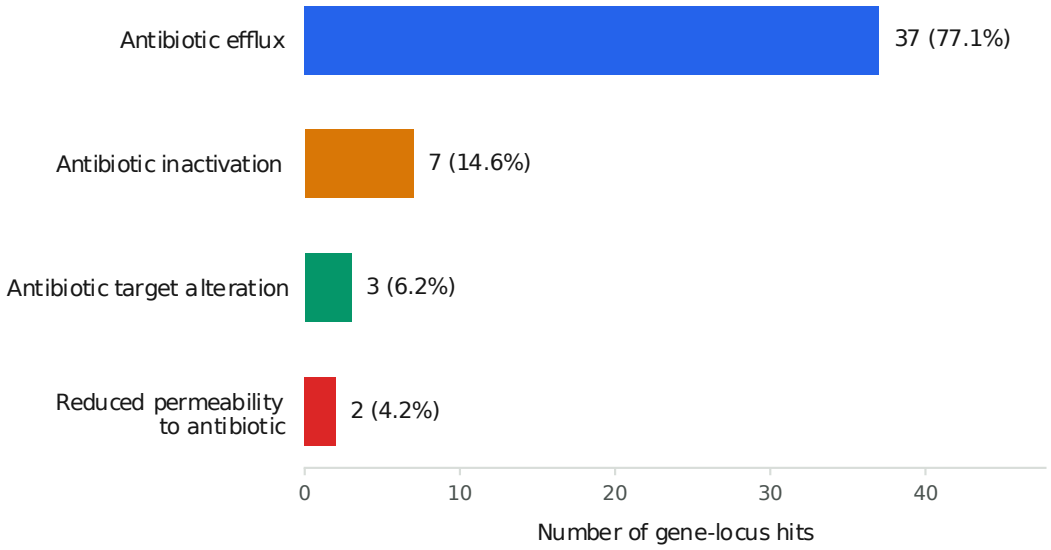



Figure 3. Distribution of antimicrobial resistance mechanisms in Escherichia coli strain S11. Bar plot showing 48 distinct CARD/ABRicate gene-locus hits retained at nucleotide identity ≥80% and reference coverage ≥90%. Antibiotic efflux (37/48), antibiotic inactivation (7/48), antibiotic target alteration (3/48) and reduced permeability to antibiotic (2/48) give 49 associations because marA contributes to two mechanisms. The denominator is 48 hits; the categories overlap. The plot includes intrinsic loci, associated regulators and four hits with partial reference coverage that passed the ≥90% threshold. Each gene-locus hit is counted once per supported mechanism.

## Experimental Design, Materials and Methods

### Sample collection and bacterial isolation

Eighty tap-water samples supplied at small food restaurants in South Dhaka, Bangladesh, were collected between December 2024 and December 2025. S11 was recovered from a sample collected on 23 December 2024 directly from the restaurant supply line, without treatment or storage at the restaurant before collection. Approximately 15 mL was collected in a sterile bottle and transported to the laboratory under refrigerated conditions. Thirty samples yielded E. coli, of which 18 were reported as ESBL positive. Two isolates, S11 and S12, were selected on the basis of their resistance phenotype, and S11 was subjected to whole-genome sequencing. Upon arrival, the samples were serially diluted up to 10^-4 in 0.85% saline solution. Spread plating was

performed on eosin methylene blue (EMB) agar from each dilution, followed by incubation for 24 hours. Colonies with a metallic green sheen and black centers were selected. Streak plating on EMB agar was repeated three times for each sample plate to ensure pure colony isolation.

Presumptive isolates were subjected to morphological and biochemical identification. Isolates were evaluated using Simmons Citrate Agar, Kligler's Iron Agar, urease test, methyl red test, Voges-Proskauer test, and Sulfur Indole Motility test. The isolate selected for genome sequencing showed biochemical characteristics consistent with E. coli: citrate negative on Simmons Citrate Agar, acidic slant and acidic butt with yellow coloration and gas production on Kligler's Iron Agar, urease negative, methyl red positive, Voges-Proskauer negative, sulfur negative, indole positive, and motility positive. S11 was selected for whole-genome sequencing and downstream genomic analyses because of its recorded resistance to multiple clinically important antimicrobial classes (Supplementary Table S7).

**Phenotypic ESBL confirmation and antimicrobial susceptibility testing**

Phenotypic ESBL production was assessed using the double-disk synergy test described by Jarlier et al. [27]. A bacterial suspension adjusted to a 0.5 McFarland standard was inoculated onto Mueller–Hinton agar. An amoxicillin/clavulanic acid disk (20/10 µg) was placed centrally, with ceftazidime (30 µg) and ceftriaxone (30 µg) disks placed 20 mm away, center to center. After incubation at 37°C for 18–24 hours, enhancement of a cephalosporin inhibition zone toward the amoxicillin/clavulanic acid disk was interpreted as a positive ESBL result.

Antimicrobial susceptibility testing was performed using the Kirby–Bauer disk diffusion method on Mueller–Hinton agar in accordance with CLSI M100 guidelines (33rd edition, 2023) [28]. The inoculum was adjusted to a 0.5 McFarland standard and plates were incubated aerobically at 37°C for 24 hours. E. coli ATCC 25922 was used as the quality-control strain. The 20-antibiotic panel and recorded susceptibility categories are provided in Supplementary Table S7. Meropenem disk content was 10 µg.

**DNA extraction, library preparation, and sequencing**

Genomic DNA was extracted using the QIAGEN DNeasy Blood & Tissue Kit, following the manufacturer's protocol. This method was used to obtain purified genomic DNA suitable for whole-genome sequencing and reduce contaminants that could interfere with library preparation. DNA concentration was measured using a Qubit fluorometer, and purity was assessed using a NanoDrop spectrophotometer. Sample S11 had a DNA concentration of 26.1 ng/µL and an A260/A230 ratio of 1.87.

Sequencing libraries were prepared as index-tagged genomic DNA libraries using 300–350 ng of purified genomic DNA. Whole-genome sequencing was performed on an Illumina NextSeq 2000 platform using sequencing-by-synthesis chemistry and paired-end 150 bp read configuration. The

sequencing report indicated 151 cycles for each read direction. The sequencing library had a concentration of 22.7 ng/µL and a fragment size of 567 bp.

**Read quality control, trimming, and genome assembly**

Read quality before and after filtering was assessed using Falco v1.3.2 on Galaxy Europe. The HRRT-cleaned reads entering fastp comprised 7,037,146 reads (3,518,573 pairs) and 1,022,578,088 bases. Adapter trimming and read filtering were performed using fastp v1.3.7. After filtering, 6,968,898 reads (3,484,449 pairs) passed quality filters, representing 99.03% of the input reads. The filtered reads contained 1,012,458,838 bases, with a mean length of 145.28 bp calculated as total bases divided by total reads, and Q20 and Q30 values of 98.22% and 92.58%, respectively [7, 8]. These counts describe the reads entering this documented filtering step. Nominal sequencing depth was 202.9×, calculated as retained bases divided by the 4,990,311-bp assembly.

The filtered paired-end reads were assembled on Galaxy Europe using Shovill v1.4.2, which wraps SPAdes v3.15.5. The recorded settings were --assembler spades, --depth 150, --minlen 200, --mincov 2 and --nocorr; Shovill trimming was disabled. The log reports an estimated input depth of 202× and subsampling to a target of 150×. The --nocorr option disabled the final Shovill correction step. Assembly quality was evaluated using QUAST v5.3.0 (Galaxy wrapper 5.3.0+galaxy2) in reference-free mode with a minimum contig length of 500 bp. The complete FASTA contained 139 contigs ≥200 bp, whereas the main QUAST statistics used 86 contigs ≥500 bp totaling 4,974,061 bp. Genome completeness and contamination were assessed using CheckM2 (Galaxy wrapper 1.1.0+galaxy1; database 1.0.2) [9, 10, 11, 12].

**Genome annotation**

Genome annotation was performed using the NCBI Prokaryotic Genome Annotation Pipeline, release 2026-06-18.build8602. The final assembly was subjected to in silico serotyping using ECTyper v2.0.0 [13]. ECTyper was used to identify O-antigen and H-antigen determinants and to assign the predicted serotype. Clermont phylogroup assignment was performed using ClermonTyping in standalone in silico PCR mode without Mash [14], using launcher release 24.02, classifier v1.4.0 and upstream commit 69da4db3cd98ab40528bacc289529323ff1407a0. BLAST+ v2.17.0+ was used with nucleotide identity ≥90%, word size 6 and classifier settings of six allowed mismatches and a perfect three-base 3′ seed. The profile was arpA−, chuA+, yjaA−, TspE4.C2− and ybgD−, giving phylogroup F. ECTyper identified wzm and wzt on contig00032 and fliC on contig00003, each with 100% identity and reference coverage (Supplementary Table S8). Plasmid replicons were identified using PlasmidFinder v2.1.6 [15], with database 2.1_2019_08_28, identity ≥95% and reference coverage ≥60%. CRISPR features were detected and visualized using CRISPRCasFinder [16] and Proksee [17]. Potential human pathogenicity was assessed using PathogenFinder2 v0.6.0 through the Center for Genomic Epidemiology web server

[18]. The final prediction was interpreted using the mean output score generated from four neural network models, using the server decision threshold of 0.5.

**Antimicrobial resistance and virulence gene analysis**

Antimicrobial resistance genes were identified from the assembled genome using ABRicate v1.4.0 [19] with multiple databases, including ResFinder [20], NCBI [21], and CARD [22], with minimum nucleotide identity and reference coverage of 80% each. The same thresholds were used for ABRicate virulence screens. The ABRicate databases had a build date of 3 April 2026. ResFinder was used to identify acquired antimicrobial resistance genes, while CARD was used to describe broader antimicrobial resistance-associated mechanisms, including efflux pumps, beta-lactamases, reduced permeability, and regulatory determinants. Virulence-associated genes were detected using ABRicate with VFDB [24] and ecoli_vf [25]. Standalone VirulenceFinder v3.2.1 [23], database 2.1.0, was run using BLASTN with minimum nucleotide identity of 90% and minimum reference coverage of 60% through the Center for Genomic Epidemiology (CGE) web server (https://genepi.dk/virulencefinder). All reported reference matches were retained in Supplementary Table S5. The assembled genome was additionally analyzed using AMRFinderPlus [26] in nucleotide-only mode. The assembled genome was analysed on Galaxy Europe using AMRFinderPlus v4.2.7 (Galaxy wrapper 4.2.7+galaxy0) and database release 2026-05-15.1 in nucleotide-only mode. Escherichia-specific mutation screening was performed with the --mutation_all option, minimum reference coverage of 0.5, curated identity thresholds and translation table 11 (Supplementary Table S9). All ABRicate and AMRFinderPlus analyses were performed using the Galaxy Europe platform (https://usegalaxy.eu).

**Visualization**

A circular genome map of Escherichia coli strain S11 was generated using Proksee [17]. Resistome visualizations were generated using custom Python scripts based on the final filtered antimicrobial resistance gene tables. The acquired resistance gene profile was visualized as a Sankey diagram linking ResFinder/ABRicate-detected genes to antimicrobial classes. Resistance mechanisms from ABRicate analysis using the CARD [19, 22] database were summarized as a bar plot, grouping determinants into antibiotic efflux, antibiotic inactivation, antibiotic target alteration and reduced permeability to antibiotic. Associated regulatory loci were included in their mapped mechanism categories.

**Limitations**

This study is limited by the analysis of a single *E. coli* genome recovered from tap water. Therefore, the results cannot be used to estimate the prevalence of ESBL-producing *E. coli* in tap water across South Dhaka. The single-isolate design limits inference regarding the dissemination pathways of ESBL-producing E. coli. Although ESBL production was phenotypically confirmed and blaCTX-M-15 was detected in the genome, other predicted resistance and virulence features are based on in silico analyses and require additional phenotypic or functional validation. The

genome is a draft assembly rather than a closed genome; therefore, the exact chromosomal or plasmid location of some genes cannot be fully resolved. Short partial virulence gene hits were not interpreted as complete functional genes. Additional isolates, long-read sequencing, plasmid reconstruction, and phenotypic virulence assays would strengthen future interpretation.

## Data Availability

Raw sequencing reads are available from NCBI SRA under SRR39239575, associated with BioProject PRJNA1478760 and BioSample SAMN61018556. The GenBank WGS project accession is JCAKBG000000000 (https://www.ncbi.nlm.nih.gov/nuccore/JCAKBG000000000). 139-contig assembly and PGAP annotation used here are supplied as supplementary data. Supplementary Tables S1–S9, including S3a and S3b, provide the curated AMRFinderPlus resistance-associated changes, database-specific AMR hits, plasmid and contig-context results, ABRicate and standalone VirulenceFinder matches, PathogenFinder2 prediction, recorded AST categories, typing summaries and AMRFinderPlus screen settings and scope. The analysis scripts are publicly available at https://github.com/rk-nishan/ecoli-s11-analysis-scripts and https://doi.org/10.6084/m9.figshare.34741515.v1.

## Ethics statement

This study used an environmental tap-water isolate and did not involve humans, animals, or human data. Therefore, ethical approval and informed consent were not required.

## Declaration of Competing Interest

The authors declare no competing financial or personal interests that could have influenced this work.


## Funding

No funding was received for this study.


## Author contributions

Shibbir Ahmed: Conceptualization, methodology, investigation, data analysis, visualization, writing – original draft.
Rasel Khan Nishan: Data curation, NCBI data submission, bioinformatics analysis, validation, visualization, writing – original draft and review.
Jannatul Ferdous[1]: Investigation, laboratory work, methodology, writing – review & editing.
Abha Rahman: Sample collection, Investigation, laboratory work, data organization, writing – review & editing.
Arafat Al Mamun: Investigation, methodology, data analysis, writing – review & editing.
Jannatul Ferdous[2]: Data analysis, validation, writing – review & editing.
Ashrafus Safa: Bioinformatics support, validation, writing – review & editing.

Md. Shariful Islam: Supervision, conceptualization, project administration, validation, writing – review & editing.
Abul Kalam Azad: Supervision, conceptualization, project administration, resources, writing – review & editing.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT (https://chatgpt.com) and Claude (https://claude.ai) for language refinement, grammar correction, clarity improvement, formatting support, and assistance in organizing manuscript sections. The authors reviewed and edited the outputs as needed. The corresponding authors, **Abul Kalam Azad** and **Md. Shariful Islam**, take full responsibility for the accuracy, integrity, and content of the manuscript.

## References


1. Kaper, J. B., Nataro, J. P., & Mobley, H. L. T. (2004). Pathogenic Escherichia coli. Nature Reviews Microbiology, 2(2), 123-140. https://doi.org/10.1038/nrmicro818
2. Paterson, D. L., & Bonomo, R. A. (2005). Extended-spectrum beta-lactamases: A clinical update. Clinical Microbiology Reviews, 18(4), 657-686. https://doi.org/10.1128/CMR.18.4.657-686.2005
3. Berendonk, T. U., Manaia, C. M., Merlin, C., Fatta-Kassinos, D., Cytryn, E., Walsh, F., Buergmann, H., Soerum, H., Norstroem, M., Pons, M.-N., Kreuzinger, N., Huovinen, P., Stefani, S., Schwartz, T., Kisand, V., Baquero, F., & Martinez, J. L. (2015). Tackling antibiotic resistance: The environmental framework. Nature Reviews Microbiology, 13(5), 310-317. https://doi.org/10.1038/nrmicro3439
4. Desye, B., Woldetsadik Mawugatie, T., Asmare, L., Tsega, Y., Melak, D., Endawkie, A., & Daba, C. (2024). Antimicrobial resistance profile of Escherichia coli in drinking water from One Health perspective in low- and middle-income countries. Frontiers in Public Health, 12, 1440908. https://doi.org/10.3389/fpubh.2024.1440908
5. Waddington, C., Carey, M. E., Boinett, C. J., Higginson, E., Veeraraghavan, B., & Baker, S. (2022). Exploiting genomics to mitigate the public health impact of antimicrobial resistance. Genome Medicine, 14(1), Article 15. https://doi.org/10.1186/s13073-022-01020-2
6. Ciufo, S., Kannan, S., Sharma, S., Badretdin, A., Clark, K., Turner, S., Brover, S., Schoch, C. L., Kimchi, A., & DiCuccio, M. (2018). Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. International Journal of Systematic and Evolutionary Microbiology, 68(7), 2386-2392. https://doi.org/10.1099/ijsem.0.002809
7. de Sena Brandine, G., & Smith, A. D. (2019). Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Research, 8, 1874. https://doi.org/10.12688/f1000research.21142.2
8. Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

9. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology, 19(5), 455-477. https://doi.org/10.1089/cmb.2012.0021
10. Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nature Methods, 20(8), 1203-1212. https://doi.org/10.1038/s41592-023-01940-w
11. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075. https://doi.org/10.1093/bioinformatics/btt086
12. Seemann, T. (n.d.-b). Shovill: Faster SPAdes assembly of Illumina reads. GitHub. Retrieved June 30, 2026, from https://github.com/tseemann/shovill
13. Bessonov, K., Laing, C., Robertson, J., Yong, I., Ziebell, K., Gannon, V. P. J., Nichani, A., Arya, G., Nash, J. H. E., & Christianson, S. (2021). ECTyper: In silico Escherichia coli serotype and species prediction from raw and assembled whole-genome sequence data. Microbial Genomics, 7(12), 000728. https://doi.org/10.1099/mgen.0.000728
14. Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., & Clermont, O. (2018). ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microbial Genomics, 4(7), e000192. https://doi.org/10.1099/mgen.0.000192
15. Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O., Villa, L., Moller Aarestrup, F., & Hasman, H. (2014). In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrobial Agents and Chemotherapy, 58(7), 3895-3903. https://doi.org/10.1128/AAC.02412-14
16. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISPRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research, 46(W1), W246-W251. https://doi.org/10.1093/nar/gky425
17. Grant, J. R., Enns, E., Marinier, E., Mandal, A., Herman, E. K., Chen, C.-Y., Graham, M., Van Domselaar, G., & Stothard, P. (2023). Proksee: In-depth characterization and visualization of bacterial genomes. Nucleic Acids Research, 51(W1), W484-W492. https://doi.org/10.1093/nar/gkad326
18. Ferrer Florensa, A., Almagro Armenteros, J. J., Kaas, R. S., Clausen, P. T. L. C., Nielsen, H., Rost, B., & Aarestrup, F. M. (2026). Whole-genome prediction of bacterial pathogenic capacity on novel bacteria using protein language models with PathogenFinder2. Bioinformatics, 42(5), btag129. https://doi.org/10.1093/bioinformatics/btag129
19. Seemann, T. (n.d.-a). ABRicate: Mass screening of contigs for antimicrobial resistance or virulence genes. GitHub. Retrieved June 30, 2026, from https://github.com/tseemann/abricate

20. Bortolaia, V., Kaas, R. F., Ruppé, E., et al. (2020). *ResFinder 4.0 for predictions of phenotypes from genotypes*. *Journal of Antimicrobial Chemotherapy, 75*(12), 3491–3500. https://doi.org/10.1093/jac/dkaa345
21. Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Connor, R., Feldgarden, M., Fine, A. M., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T. D., O'Sullivan, C., Schmieder, E., Skripchenko, Y., et al. (2025). Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Research, 53(D1), D20–D29. https://doi.org/10.1093/nar/gkae979
22. Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., ... McArthur, A. G. (2020). CARD 2020: Antibiotic resistome surveillance with the Comprehensive Antibiotic Resistance Database. Nucleic Acids Research, 48(D1), D517-D525. https://doi.org/10.1093/nar/gkz935
23. Joensen, Katrine G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., & Aarestrup, F. M. (2014). Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *Journal of Clinical Microbiology, 52*(5), 1501–1510. https://doi.org/10.1128/JCM.03617-13
24. 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 Research, 44*(D1), D694–D697. https://doi.org/10.1093/nar/gkv1239
25. Public Health Agency of Canada. (n.d.). *ecoli_vf: Curated virulence factors for Escherichia coli*. https://github.com/phac-nml/ecoli_vf
26. Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H., & Klimke, W. (2021). *AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence*. *Scientific Reports, 11*, 12728. https://doi.org/10.1038/s41598-021-91456-0
27. Jarlier, V., Nicolas, M. H., Fournier, G., & Philippon, A. (1988). Extended broad-spectrum beta-lactamases conferring transferable resistance to newer beta-lactam agents in Enterobacteriaceae: Hospital prevalence and susceptibility patterns. Reviews of Infectious Diseases, 10(4), 867–878. https://doi.org/10.1093/clinids/10.4.867
28. Clinical and Laboratory Standards Institute. (2023). Performance Standards for Antimicrobial Susceptibility Testing. 33rd ed. CLSI supplement M100. Clinical and Laboratory Standards Institute.

## Supplementary Files

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

- S11supplementaryfiles.zip