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.

Phylogroup-Driven Niche Adaptation and Virulence in Escherichia coli

Di Dong
Department of Laboratory Medicine, Nanjing Drum Tower Hospital Clinical College of Nanjing University of Chinese Medicine

Peitao Jiang
Department of Clinical Laboratory, Yangzhou Yizheng Hospital

Hui Xie
Department of Clinical Laboratory, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University

Zhengnan Yang
Department of Clinical Laboratory, Yizheng Hospital of Nanjing Drum Tower Hospital Group

Chang Liu
Department of Laboratory Medicine, Nanjing Drum Tower Hospital Clinical College of Nanjing University of Chinese Medicine

Xiaoli Cao
cao-xiao-li@163.com

Department of Laboratory Medicine, Nanjing Drum Tower Hospital Clinical College of Nanjing University of Chinese Medicine

Research Article

Keywords: Escherichia coli, Phylogroup, Multi-Locus Sequence Typing, Virulence Factors, Population Structure

Posted Date: May 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9289690/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background:

Escherichia coli encompasses diverse strains ranging from commensals to major human pathogens. How this diversity aligns with phylogenetic background, ecological niche, and virulence potential remains a central question for genomic surveillance.

Methods:

We analyzed 4,832 high-quality *E. coli* genomes from public databases and our own isolates. Phylogroups were assigned via *in silico* PCR, and available metadata on host and geography were integrated. Population structure was resolved using both Achtman and Pasteur MLST schemes. Virulence factors were screened against the Virulence Factors Database (VFDB).

Results:

Phylogroups A (32.5%), B1 (21.5%), and B2 (18.8%) predominated, each with distinct niche preferences: B2 was largely human-adapted, whereas A and B1 behaved as generalists. Geographical patterns were evident in the metadata but appeared heavily influenced by sampling disparities across regions and host types. The Achtman and Pasteur MLST schemes yielded different numbers of sequence types (STs) (805 vs. 570) yet showed strong concordance at the phylogroup level. Virulence profiles followed phylogroup-specific patterns—for instance, extra-intestinal pathogenic *E. coli* (ExPEC) factors concentrated in B2, while intestinal pathogenicity genes were enriched in phylogroup E. Lineages such as ST131 (Achtman) / ST43 (Pasteur) maintained conserved virulence repertoires that define successful clones.

Conclusions:

Our findings reinforce the model of phylogroup-driven niche adaptation in *E. coli*, with distinct virulence gene repertoires shaping ecological success. Public genomic datasets offer powerful tools for uncovering host-associated patterns, but their inherent sampling biases—particularly in geographical coverage—warrant caution in interpretation. This integrated analysis provides a framework for dissecting lineage-specific pathogenicity and underscores the need for balanced sampling in future surveillance efforts.

Introduction

Escherichia coli (E. coli) presents a paradox in clinical microbiology: a commensal of the human and animal gut, yet also a pathogen capable of causing everything from self-limiting diarrhea to life-

threatening sepsis and meningitis[1, 2]. This duality is rooted in its genomic plasticity—different strains carry distinct repertoires of virulence genes that dictate where they can colonize and what disease they can cause [3].

Pathogenic *E. coli* are traditionally classified into pathotypes based on clinical presentation and virulence mechanisms. Intestinal pathogens (InPEC) include enteropathogenic *E. coli* (EPEC), which causes infantile diarrhea via attaching and effacing lesions [4]; enterotoxigenic *E. coli* (ETEC), the agent of traveler's diarrhea[5]; enterohemorrhagic *E. coli* (EHEC), associated with Shiga toxin-mediated hemolytic uremic syndrome[6]; enteroinvasive *E. coli* (EIEC), which produces a dysentery-like illness [7]; and enteroaggregative *E. coli* (EAEC), characterized by a "stacked-brick" biofilm on the intestinal mucosa[8]. Beyond the gut, extraintestinal pathogenic *E. coli* (ExPEC)—including uropathogenic *E. coli* (UPEC) and neonatal meningitis-associated *E. coli* (NMEC)—carry VFs that enable colonization of sterile sites[9].

A key organizing principle for *E. coli* diversity is the phylogroup classification, which assigns strains to major lineages such as A, B1, B2, C, D, E, F, and several cryptic clades [10, 11]. These phylogroups are not merely taxonomic labels; they predict ecological niche and pathogenic potential. Commensal strains cluster predominantly in phylogroups A and B1, whereas ExPEC strains are heavily enriched in B2 and D[12]. Intestinal pathotypes also show strong phylogroup affinities: typical EPEC is often found in phylogroup C, EHEC (including O157:H7) in phylogroup E, and ETEC in A and B1[13, 14]. These associations reflect a deeper evolutionary reality—the core genome backbone constrains or enables the acquisition of specific virulence traits[15–17].

At a finer scale, multilocus sequence typing (MLST) has become the standard for tracking *E. coli* clones. Two schemes predominate: the Achtman scheme (7 housekeeping genes) and the Pasteur scheme (8 genes). Both have been instrumental in identifying high-risk lineages, most notably ST131 (Achtman), a globally disseminated ExPEC clone within phylogroup B2 [18]. Yet the two schemes use different loci and databases, raising questions about whether they yield comparable views of population structure. A systematic, large-scale comparison of their concordance—particularly in linking sequence types (STs) to phylogroups and virulence profiles—has been lacking.

Existing studies have explored the phylogroup-virulence connection, but typically at a coarse level (phylogroups) or for a handful of prominent STs. What is missing is an integrated, genomic-scale analysis that brings together phylogroup assignment, both MLST schemes, and detailed VF profiles across thousands of genomes. Such an approach would allow us to ask: Do the Achtman and Pasteur schemes converge on the same phylogenetic and pathogenic framework? Can we define ST-specific virulence signatures that are robust across typing schemes? And how is the distribution of VFs structured across the entire *E. coli* population?

To address these questions, we analyzed 4,899 *E. coli* genomes from public databases and our own isolates. We aimed to: (1) map phylogroup distribution across this large collection; (2) compare the two

MLST schemes for their resolution and concordance; (3) profile VF distribution across phylogroups; and (4) identify lineage-specific virulence signatures at the ST level. Our findings offer an integrated view of *E. coli* population structure, revealing that successful clones carry conserved, lineage-specific virulence repertoires that are captured consistently regardless of the typing scheme used.

Materials and methods

1. Genome Sequence Acquisition and Processing

A total of 4,900 *E. coli* genome assemblies were analyzed, including 137 in-house sequenced isolates from a multicenter clinical surveillance study in Nanjing (2015–2022; BioProject PRJNA989674) and 4,763 publicly available complete genomes from NCBI (accessed September 15, 2024). Public genomes spanned isolation dates from 1884 to 2024, though sequencing and deposition occurred in recent years.

GenBank files were downloaded using the NCBI datasets command-line tool and parsed with custom Perl scripts. Gene prediction and annotation were performed uniformly with Prokka v1.14.5[19]. Genome quality was assessed using CheckM v1.1.3[20] and QUAST v5.0.2[21]. Genomes were considered high-quality if they met: completeness > 90%, contamination < 5%, contig count ≤ 500, and N50 ≥ 40 kb. Of the 4,900 assemblies, 4,832 passed these criteria and were retained for further analysis.

2. Phylogroup Determination

Phylogroups were assigned via *in silico* PCR using primers targeting 12 lineage-specific markers: the 7 primer pairs for major phylogroups (A, B1, B2, C, D, E, F) from Clermont et al. 2013[11] and the 5 primer pairs for Clades I–V from Clermont et al. 2011[22] (Supplementary Table S1). Primer binding sites were identified using a custom script requiring ≥ 95% identity, ≥ 65% query coverage, and a perfect match at the 3' terminal base. Amplicon sequences were validated by BLASTn (BLAST+ v2.13.0) against the primer sequences. Final phylogroup assignment followed the decision algorithm of Clermont et al. 2013 [11], with Clades I–V classified according to Clermont et al. 2011[22].

3. Metadata Integration

Host origin, country of isolation, and previously assigned phylogroup designations were extracted from GenBank files using custom Perl scripts. These data were consolidated into a structured dataset to assess ecological and geographical distribution patterns, focusing on host specificity and global dissemination.

4. MLST Typing

MLST was performed using both the Achtman (7-gene) and Pasteur (8-gene) schemes. Alleles were identified by BLAST+ against reference allele libraries. For the Achtman scheme, the seven housekeeping genes were *adk, fumC, gyrB, icd, mdh, purA, recA*; for the Pasteur scheme: *dinB, icdA, pabB, polB, putP, trpA, trpB, uidA*. A known ST was assigned when the allele profile matched an existing entry; novel allele combinations or alleles were designated as novel STs.

5. Virulence Gene Profiling

VFs were identified in the 4,832 high-quality genomes (plus 33 low-quality genomes that contained virulence genes, see Table S2). One complete genome (GCA_022493475.1) with a total length < 1 Mb was excluded. Annotations from Prokka were used as input. The Virulence Factors Database (VFDB, nucleotide format, accessed September 13, 2024) was obtained from http://www.mgc.ac.cn/VFs/download.htm[23]. Coding sequences (CDS) were extracted from each genome and aligned against VFDB using BLASTn with the following criteria: E-value ≤1e − 5, sequence identity ≥ 70%, alignment coverage ≥ 90%, and alignment length ≥ 30 bp. Of the 4,899 genomes analyzed, 4,865 carried at least one detectable virulence gene. The remaining 34 low-quality genomes (all from public sources) yielded no hits under these criteria.

Results

1. Phylogroup Distribution

Among the 4,832 high-quality *E. coli* genomes, phylogroup A was most prevalent (1,572 isolates, 32.5%), followed by B1 (1,038, 21.5%) and B2 (907, 18.8%). Phylogroups D (428, 8.9%), E (443, 9.2%), F (209, 4.3%), and C (190, 3.9%) were less frequent. Seven isolates fell into Clade I, and 32 (0.7%) could not be assigned to any known phylogroup (Figure 1).

2. **Host and Geographic Associations**

Host metadata revealed distinct patterns: phylogroup B2 isolates were predominantly human-derived (49.3% from *Homo sapiens*), whereas phylogroups A and B1 were recovered from a broader range of hosts, including humans, livestock, and poultry. Phylogroup D was prevalent in both human and swine isolates, and phylogroup F was frequently isolated from human and avian hosts (Table 1).

Geographic patterns were heavily influenced by sampling bias. Asia contributed the largest number of genomes for most lineages; for example, a high proportion of phylogroup B2 genomes originated from South Korea, while phylogroup E genomes were predominantly from the USA and UK. However, host composition varied markedly across regions, and a substantial portion of genomes lacked host or location metadata (e.g., 373, 237, and 187 genomes for phylogroups A, B1, and B2, respectively) (Figure 2).

3. **MLST Diversity**

MLST typing using the Achtman scheme identified 805 sequence types (STs). The most common were ST10 (464 isolates, 9.6%), ST131 (434, 8.9%), ST11 (337, 6.9%), ST167 (124, 2.5%), ST410 (124, 2.5%), and ST38 (120, 2.5%). Forty-seven novel STs were identified, and 122 isolates could not be assigned to a known type.

The Pasteur scheme yielded 570 STs, with ST2 (316, 6.5%), ST43 (308, 6.4%), ST262 (204, 4.2%), and ST8 (117, 2.4%) predominating. A substantially higher proportion of isolates were untypeable under this scheme: 27.8% (1,341 isolates) were unknown and 6.0% (292) were novel (Table 2, Figure 3).

5. Comparison of Achtman and Pasteur MLST Schemes

The two MLST schemes differed in resolution but showed strong concordance at the phylogroup level. The Pasteur scheme had higher discriminatory power, identifying fewer STs per phylogroup (e.g., phylogroup A: 132 vs. 282 in Achtman) but generating more novel and unclassifiable STs (phylogroup A: 123 novel and 452 unknown vs. 16 and 40 for Achtman). Consequently, the predominant STs within each phylogroup differed between schemes—for example, ST131 (Achtman) versus ST43 (Pasteur) in phylogroup B2.

Despite these differences, dominant STs from both schemes showed strong phylogroup specificity. ST10 and ST131 (Achtman) were exclusively associated with phylogroups A and B2, respectively; similarly, ST2 and ST43 (Pasteur) mapped to the same phylogroups (Table 3).

6. Virulence Gene Distribution

Virulence genes were categorized into 11 functional classes. Adhesins, effector delivery systems, mobile genetic elements, and nutrition/metabolic factors were the most abundant categories; competitive factors, biofilm formation, exoenzymes, exotoxins, immunomodulators, invasins, and regulators were less prevalent.

5.1 **Phylogroup-Level Virulence Patterns**

Prevalence of key VFs varied by phylogroup (Table 4-1). The biofilm-associated gene (VFC0271) was most common in phylogroups B2 (87.1%) and E (85.8%). The exoenzyme gene (VFC0251) was concentrated in phylogroup E (76.1%), with minimal presence elsewhere (0–6.3%). Exotoxin genes (VFC0235) were prevalent across most phylogroups, exceeding 95% in D, E, B2, and F, but dropped to 40.5% in phylogroup C.

5.2 ST-Level Virulence Signatures

Predominant STs from both schemes exhibited distinct virulence profiles that aligned with their phylogroup assignments (Tables 4-2, 4-3).

Under the Achtman scheme, ST10 (phylogroup A) showed high prevalence of exoenzyme (96.1%) and biofilm (79.3%) genes but low exotoxin carriage. In contrast, ST131, ST11, and ST38—all belonging to

phylogroup B2 or related lineages—displayed a “high-virulence” profile with near-ubiquitous exotoxin and biofilm genes. ST167 was notable for its high carriage rate of the “Others” virulence category (92.7%), while ST410 showed comparatively lower virulence gene prevalence overall.

Under the Pasteur scheme, ST43 (phylogroup B2) carried classic ExPEC factors, including fimbrial clusters (≥98.7%), yersiniabactin (100%), and the toxin Sat (55.8%). ST8 (also phylogroup B2) was defined by Afa/Dr adhesins (*afaB*, 46.6%) and the ETT2 secretion system. ST2 (phylogroup A) possessed a unique K-locus and a Type VI secretion system (T6SS), whereas ST262 showed minimal VF content. Notably, a biofilm-associated gene was conserved across most dominant STs in both schemes—for example, 97.4% in ST43 and over 95.5% in ST131.

5.3. Phylogroup-Specific Virulence Markers

Certain VFs showed strong phylogroup specificity, serving as distinctive markers for individual lineages (Table S3). Phylogroup B2 was the primary reservoir for classic ExPEC-associated factors. These included P fimbriae (*papA*, 51.8%), S fimbriae (*sfaG*, 16.6%), the toxin Cnf1 (26.6%), and the genotoxic *pks* island (*clbB*, 25.6%). A Type VI secretion system (T6SS) was also highly enriched in this phylogroup, with its core component *vgrG* present in 90.5% of B2 isolates. In contrast, the CFA/I fimbrial cluster (*cfaA*) was characteristic of phylogroups B1 and C, with carriage rates of 96.0% and 97.9%, respectively. Phylogroup E was defined by intestinal pathogenesis genes, including the LEE pathogenicity island (*eae*, 80.1%) and Shiga toxin (*stx2A*, 98.6%). Phylogroup F possessed a unique fimbrial cluster (ECSMS35_RS19770) that was largely absent in other phylogroups, present in 28.7% of F isolates.

5.4 Lineage-Specific Virulence Repertoires

Beyond phylogroup-level patterns, individual STs carried distinct VF assemblages that further defined their pathogenic niches (Tables S4, S5).

Among Achtman STs, ST131 was characterized by yersiniabactin (*fyuA*, 99.5%), a specific fimbrial cluster (98.4%), and the Sat toxin (80.7%). ST11 carried the LEE island (*tir*, 100%), Shiga toxin (*stx2A*, 92.9%), and hemolysin (*hlyA*, 95.5%). ST38 displayed a hybrid profile combining Afa/Dr adhesins with a complete ETT2 system (*eivF*, 94.2%). ST167 was defined by a dominant T6SS (*tssL*, 100%). Intra-lineage specialization was also evident within the ST10 complex, where one sub-clone carried CFA/I fimbriae (100%) and another carried *fimK* (100%), suggesting functional diversification within a single ST.

The Pasteur scheme revealed complementary patterns. ST43 mirrored ST131, with high carriage of fimbrial clusters (98.7%), yersiniabactin (100%), and toxins Cnf1 (31.0%) and Sat (55.8%). ST8 was characterized by Afa/Dr adhesins (*afaB*, 46.6%) and ETT2 (*eivF*, 95.8%). ST2 carried a unique K-locus and T6SS, while ST262 lacked most canonical virulence genes.

Discussion

This study provides a genomic survey of *E. coli* population structure, ecological distribution, and virulence gene repertoires across nearly 5,000 high-quality genomes. Our findings reinforce established models of phylogroup-driven niche adaptation while revealing finer-scale phylogenetic and pathogenic specializations.

The prevalence of phylogroups A, B1, and B2 in our collection mirrors global trends reported in previous studies[11, 24]. The host association patterns we observed—phylogroup B2 predominantly in human samples, phylogroups A and B1 across a broader range of hosts—support the concept of specialist versus generalist ecotypes within the species[25] [26]. Phylogroup D was frequently recovered from swine and human hosts, and phylogroup F from avian and human sources, suggesting zoonotic potential for these lineages[27, 28].

Geographic patterns must be interpreted with caution. In our dataset, Asia contributed the largest number of genomes, and certain lineages appeared concentrated in specific countries (e.g., phylogroup B2 in South Korea, phylogroup E in the USA and UK). However, these patterns are confounded by uneven global sampling and disparities in host representation across regions. Public databases systematically over-represent human clinical isolates from high-surveillance regions while under-sampling environmental and livestock reservoirs elsewhere [29]. Any apparent “geographical distribution” reported in meta-analyses of this type likely reflects database artifacts as much as—or more than—authentic microbial ecology, a challenge noted in other genomic surveillance studies[30]. Balanced, prospective sampling designs will be essential for true phylogeographic inference.

The two MLST schemes differed in resolution but showed strong concordance at the phylogroup level. The Pasteur scheme had higher discriminatory power, yielding more novel and untypable STs, yet both schemes assigned strains to major phylogenetic lineages with similar overall patterns. This concordance likely reflects their shared ability to distinguish deep-branching phylogroups (e.g., B2 versus A) that form the primary phylogenetic framework of *E. coli*, rather than indicating equivalent resolution at finer taxonomic scales. Practically, this finding suggests that for large-scale studies focused on broad population structure, the two schemes are largely interchangeable—a point supported by prior methodological comparisons[28, 31]. The consistent mapping of dominant STs across schemes—ST10 (Achtman) to ST2 (Pasteur) in phylogroup A, and ST131 (Achtman) to ST43 (Pasteur) in phylogroup B2—further confirms that successful epidemic clones are tightly linked to their core phylogenetic background[32]. These cross-scheme correlations reinforce the clonal model of *E. coli* evolution[33] and demonstrate that both schemes converge on the same high-impact lineages, even if their finer-scale substructure classifications differ.

Our VF analysis revealed highly specific, phylogroup-associated signatures. Phylogroup B2 was the primary reservoir for classic ExPEC factors—P fimbriae, Cnf1, and the *pks* island—consistent with its established role in extra-intestinal infections[34]. Conversely, the CFA/I fimbrial cluster was strongly associated with phylogroups B1 and C, and the LEE pathogenicity island with phylogroup E, aligning with

their known roles in intestinal pathogenesis[35, 36]. The enrichment of a Type VI secretion system (T6SS) in phylogroup B2 is of particular interest. T6SS has been implicated in inter-bacterial competition and host adaptation in *E. coli*, with recent studies suggesting roles in biofilm formation and colonization[37]. Whether this system contributes specifically to the success of B2-associated ExPEC lineages warrants further investigation.

Integrating MLST with virulence profiling revealed that successful clones carry distinct molecular signatures that transcend the typing scheme used. ST131 (Achtman) was defined by yersiniabactin and a specific fimbrial cluster; ST43 (Pasteur), its counterpart in phylogroup B2, showed an almost identical repertoire. ST11 carried the LEE island, Shiga toxin, and hemolysin—a combination that explains its association with severe intestinal disease[38]. Within the ST10 complex, we identified sub-lineages differentiated by CFA/I fimbriae versus *fimK*, illustrating how integrated genomic approaches can uncover micro-evolutionary specializations within broad lineages[39, 40]. A biofilm-associated gene was conserved across most dominant STs in both schemes: over 95% in ST131 and 97% in ST43, suggesting that biofilm formation may be a core, near-essential virulence trait for successful *E. coli* clones. This aligns with growing evidence linking biofilm formation to colonization efficiency and antimicrobial tolerance[41].

Several limitations should be considered. First, the reliance on public genomic databases introduces systematic biases. Host representation is skewed toward human clinical isolates, and geographic metadata are incomplete or unevenly distributed. As a result, the patterns of “geographic distribution” and “host specificity” reported here likely reflect research and surveillance priorities as much as—or more than—the true ecological landscape of *E. coli*. Second, our *in silico* approaches for phylogroup assignment and virulence gene detection, while stringent, are based on sequence similarity. They may fail to capture highly divergent gene families and cannot infer functional activity or regulatory context. Third, this study identifies correlations at the genomic level but does not establish causal mechanisms. The associations reported here will require functional validation, for example through isogenic mutant models or experimental infection studies.

Conclusion

In summary, this large-scale genomic analysis reinforces the paradigm of phylogroup-specific niche adaptation in *E. coli*, with ecological specialization underpinned by conserved, lineage-specific VF repertoires. These genetic signatures define the pathogenic potential of globally successful clones such as ST131. Our findings also highlight a critical methodological point: while public genomic resources are invaluable for uncovering host-associated patterns, inferring true geographic or macro-ecological distributions from them requires caution due to inherent sampling biases. This study provides a phylogeny-aware framework for risk assessment of emerging strains and underscores the need for prospective, balanced sampling designs in future surveillance efforts. Functional validation of lineage-

specific virulence markers and longitudinal, community-structured surveillance will be essential next steps.

Declarations

Declarations

None

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from GenBank and the accession number and web link to datasets for the provided name of these strains were shown in table S2.xlsx.

Competing interests

The authors declare that they have no competing interests

Funding

This work was supported by the National Natural Science Foundation of China (81902124).

Author Contribution

Di Dong and Peitao Jiang were responsible for the main data collection and statistical analysis. Hui Xie and Zhengnan Yang participated in the preparation of data verification. Chang Liu and Xiaoli Cao were in charge of experimental design, project conception and research guidance, as well as providing funding support and project management. All authors participated in the drafting and revision of the manuscript, reviewed and approved the final version for publication.

Data Availability

The datasets used and/or analyzed during the current study are available from GenBank and the accession number and web link to datasets for the provided name of these strains were shown in table

S2.xlsx.

References

1. Gomes TAT, Elias WP, Scaletsky ICA, Guth BEC, Rodrigues JF, Piazza RMF, Ferreira LCS, Martinez MB. Diarrheagenic Escherichia coli[J]. Brazilian J Microbiol. 2016;47:3–30.
2. Riley LW, Blanton RE. Distinguishing Pathovars from Nonpathovars: Escherichia coli[J]. Microbiol Spectr. 2020, 8.
3. Geurtsen J, de Been M, Weerdenburg E, Zomer A, McNally A, Poolman J. Genomics and pathotypes of the many faces ofEscherichia coli[J]. FEMS Microbiol Rev. 2022, 46.
4. Kaur P, Dudeja PK. Pathophysiology of Enteropathogenic Escherichia coli-induced Diarrhea[J]. Newborn. 2023;2:102–13.
5. Mirhoseini A, Amani J, Nazarian S. Review on pathogenicity mechanism of enterotoxigenic Escherichia coli and vaccines against it[J]. Microb Pathog. 2018;117:162–9.
6. Nguyen Y, Sperandio V, Enterohemorrhagic E. coli (EHEC) pathogenesis[J]. Front Cell Infect Microbiol. 2012, 2.
7. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, Morabito S, Prosseda G. The Intriguing Evolutionary Journey of Enteroinvasive E. coli (EIEC) toward Pathogenicity[J]. Front Microbiol. 2017, 8.
8. Hebbelstrup Jensen B, Olsen KEP, Struve C, Krogfelt KA, Petersen AM. Epidemiology and Clinical Manifestations of Enteroaggregative Escherichia coli[J]. Clin Microbiol Rev. 2014;27:614–30.
9. Sora VM, Meroni G, Martino PA, Soggiu A, Bonizzi L, Zecconi A. Extraintestinal Pathogenic Escherichia coli: Virulence Factors and Antibiotic Resistance[J]. Pathogens. 2021, 10.
10. O Clermont SB, Bingen E. Rapid and Simple Determination of the Escherichia coli Phylogenetic Group[J]. Appl Environ Microbiol. 2000;Oct(10):4555–8.
11. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups[J]. Environ Microbiol Rep. 2013;5:58–65.
12. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli[J]. Nat Rev Microbiol. 2010;8:207–17.
13. Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, Rasko DA, Payne SM. A Comparative Genomic Analysis of Diverse Clonal Types of EnterotoxigenicEscherichia coliReveals Pathovar-Specific Conservation[J]. Infect Immun. 2011;79:950–60.
14. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent Advances in Understanding Enteric Pathogenic Escherichia coli[J]. Clin Microbiol Rev. 2013;26:822–80.
15. Bugarel M, Martin A, Fach P, Beutin L. Virulence gene profiling of enterohemorrhagic (EHEC) and enteropathogenic (EPEC) Escherichia coli strains: a basis for molecular risk assessment of typical and atypical EPEC strains[J]. BMC Microbiol. 2011, 11.

16. Clermont O, Condamine B, Dion S, Gordon DM, Denamur E. The E phylogroup ofEscherichia coliis highly diverse and mimics the wholeE. colispecies population structure[J]. Environ Microbiol. 2021;23:7139–51.
17. Alfinete NW, Bolukaoto JY, Heine L, Potgieter N, Barnard TG. Virulence and phylogenetic analysis of enteric pathogenic Escherichia coli isolated from children with diarrhoea in South Africa[J]. Int J Infect Dis. 2022;114:226–32.
18. Nicolas-Chanoine MH, Blanco J, Leflon-Guibout V, Demarty R, Alonso MP, Canica MM, Park YJ, Lavigne JP, Pitout J, Johnson JR. Intercontinental emergence of Escherichia coli clone O25:H4-ST131 producing CTX-M-15[J]. J Antimicrob Chemother. 2007;61:273–81.
19. Seemann T. Prokka: rapid prokaryotic genome annotation[J]. Bioinformatics. 2014;30:2068–9.
20. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes[J]. Genome Res. 2015;25:1043–55.
21. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies[J]. Bioinformatics. 2013;29:1072–5.
22. Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. Characterization of the cryptic Escherichia lineages: rapid identification and prevalence[J]. Environ Microbiol. 2011;13:2468–77.
23. Liu B, Zheng D, Zhou S, Chen L, Yang J. VFDB 2022: a general classification scheme for bacterial virulence factors[J]. Nucleic Acids Res. 2022;50:D912–7.
24. Escobar-Páramo P, Le Menac'h A, Le Gall T, Amorin C, Gouriou S, Picard B, Skurnik D, Denamur E. Identification of forces shaping the commensal Escherichia coli genetic structure by comparing animal and human isolates[J]. Environ Microbiol. 2006;8:1975–84.
25. Condamine B, Morel-Journel T, Tesson F, Royer G, Magnan M, Bernheim A, Denamur E, Blanquart F, Clermont O. Strain phylogroup and environmental constraints shape Escherichia coli dynamics and diversity over a 20-year human gut time series[J]. ISME J. 2025, 19.
26. Silva N, Igrejas G, Gonçalves A, Poeta P. Commensal gut bacteria: distribution of Enterococcus species and prevalence of Escherichia coli phylogenetic groups in animals and humans in Portugal[J]. Ann Microbiol. 2011;62:449–59.
27. Cortés P, Blanc V, Mora A, Dahbi G, Blanco JE, Blanco M, López C, Andreu A, Navarro F, Alonso MP, et al. Isolation and Characterization of Potentially Pathogenic Antimicrobial-Resistant Escherichia coli Strains from Chicken and Pig Farms in Spain[J]. Appl Environ Microbiol. 2010;76:2799–805.
28. Wang Y, Zhou J, Li X, Ma L, Cao X, Hu W, Zhao L, Jing W, Lan X, Li Y, et al. Genetic diversity, antimicrobial resistance and extended-spectrum β-lactamase type of Escherichia coli isolates from chicken, dog, pig and yak in Gansu and Qinghai Provinces, China[J]. J Global Antimicrob Resist. 2020;22:726–32.
29. Pursey E, Dimitriu T, Gaze WH, Westra ER, van Houte S. The distribution of antimicrobial resistance genes across phylogroup, host species and geography in 16,000 publicly-available E. coli genomes[J]. medRxiv. 2023, July 12, 2023.

30. Johnston BD, Thuras P, Porter SB, Anacker M, VonBank B, Vagnone PS, Witwer M, Castanheira M, Johnson JR. Global molecular epidemiology of carbapenem-resistant Escherichia coli (2002–2017) [J]. European Journal of Clinical Microbiology & Infectious Diseases; 2021.
31. Gaiarsa S, Batisti Biffignandi G, Esposito EP, Castelli M, Jolley KA, Brisse S, Sassera D, Zarrilli R. Comparative Analysis of the Two Acinetobacter baumannii Multilocus Sequence Typing (MLST) Schemes[J]. Front Microbiol. 2019, 10.
32. Decano AG, Downing T. An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates[J]. Sci Rep. 2019, 9.
33. Pitout JDD, Finn TJ. The evolutionary puzzle of Escherichia coli ST131[J]. Infect Genet Evol. 2020, 81.
34. Yun CS, Moon B-Y, Hwang M-H, Lee S-K, Ku B-K, Lee K. Characterization of the pathogenicity of extraintestinal pathogenic Escherichia coli isolates from pneumonia-infected lung samples of dogs and cats in South Korea[J]. Sci Rep. 2023, 13.
35. Yang X, Sun H, Fan R, Fu S, Zhang J, Matussek A, Xiong Y, Bai X. Genetic diversity of the intimin gene (eae) in non-O157 Shiga toxin-producing Escherichia coli strains in China[J]. Sci Rep. 2020, 10.
36. Sébastien O, Leclercq MB, David GE, Smith P. Germon. Lipopolysaccharide core type diversity in the Escherichia coli species in association with phylogeny, virulence gene repertoire and distribution of type VI secretion systems[J]. Microb Genom. 2021 Sep: 7(9):000652.
37. Nesporova K, Steemans B, Govers SK, Ozer EA. Large-scale genomic analysis reveals the distribution and diversity of type VI secretion systems in Escherichia coli[J]. mSystems. 2025, 10.
38. Ruzickova M, Karola I, Nohejl T, Sukkar I, Palkovicova J, Papousek I, Cummins ML, Djordjevic SP, Dolejska M. Comparative genomics of multi-drug resistant Escherichia coli ST131 from human, animal, and environmental sources in the Czech Republic[J]. Ecotoxicology and Environmental Safety; 2025. p. 299.
39. Machado MAM, Panzenhagen P, Lázaro C, Rojas M, Figueiredo EES, Conte-Junior CA. Unveiling the High Diversity of Clones and Antimicrobial Resistance Genes in Escherichia coli Originating from ST10 across Different Ecological Niches[J]. Antibiotics. 2024, 13.
40. Shoaib M, Ali MM, Tang M, Weiwei W, Zhang X, He Z, Wu Z, Wang S, Hao B, Li R, Pu W. Genomic and phylogeographical analysis revealed CTX-M-55 producing Escherichia coli ST10 and ST2325 clones of One Health concern from dairy farm waste in Gansu, China[J]. One Health. 2025, 20.
41. Karigoudar RM, Karigoudar MH, Wavare SM, Mangalgi SS. Detection of biofilm among uropathogenic Escherichia coli and its correlation with antibiotic resistance pattern[J]. J Lab Physicians. 2020;11:017–22.

Tables

Tables 1 to 4 are available in the Supplementary Files section.

Figures

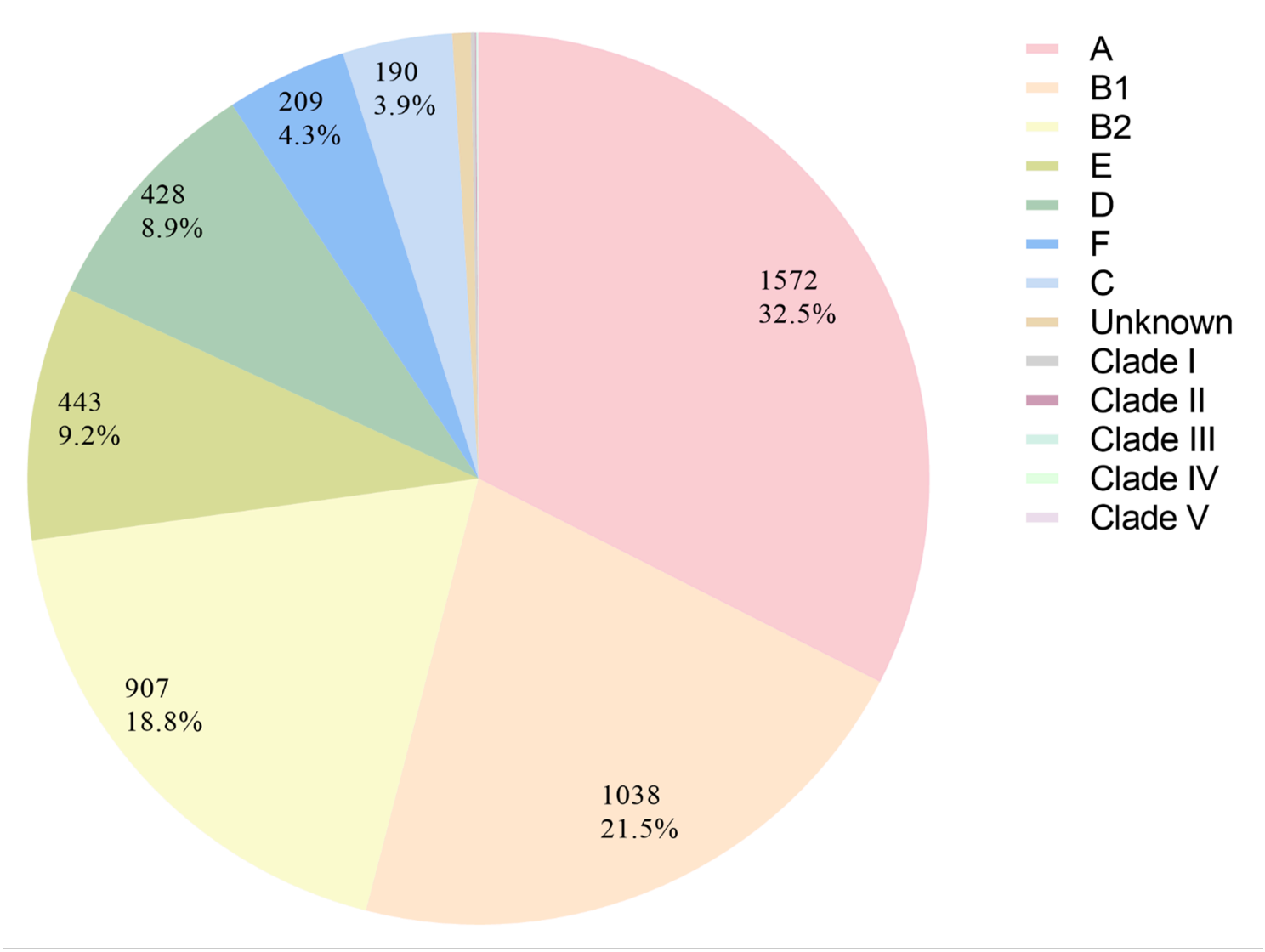


Figure 1

The proportion of phylogenetic groups in the *E. coli* genome classification. The pie chart illustrates the relative abundance of major *E. coli* phylogroups (A, B1, B2, C, D, E, F), along with an "Unknown" category for isolates that could not be classified. Additionally, the proportions for Clades I through V are indicated.

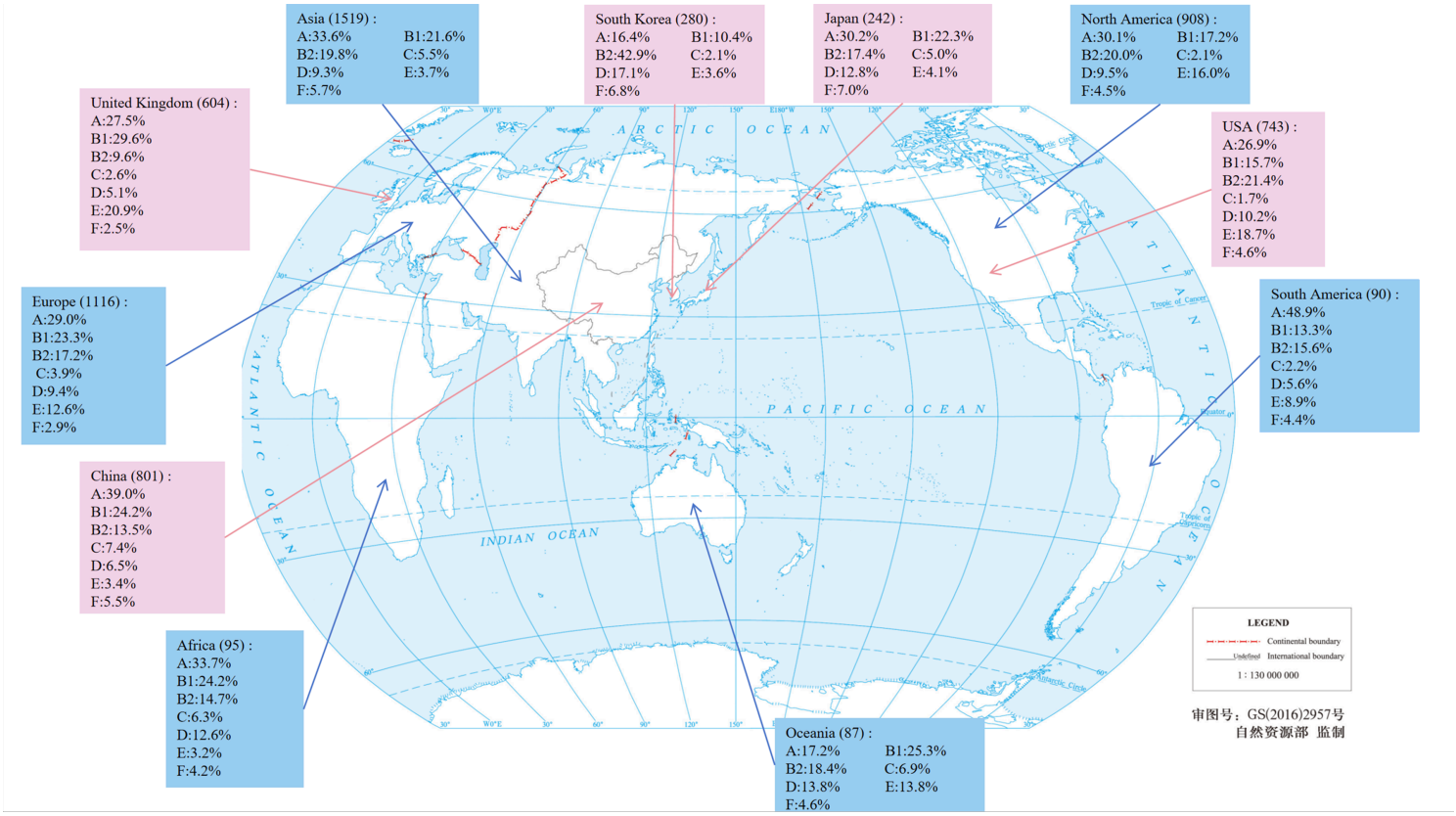


Figure 2

Geographical Distribution Patterns of *E. coli* Phylogroups. The proportional abundance of phylogroups (A, B1, C, D, E, F) among *E. coli* isolates from different countries and regions. The total number of isolates for each region is indicated in parentheses following its name.

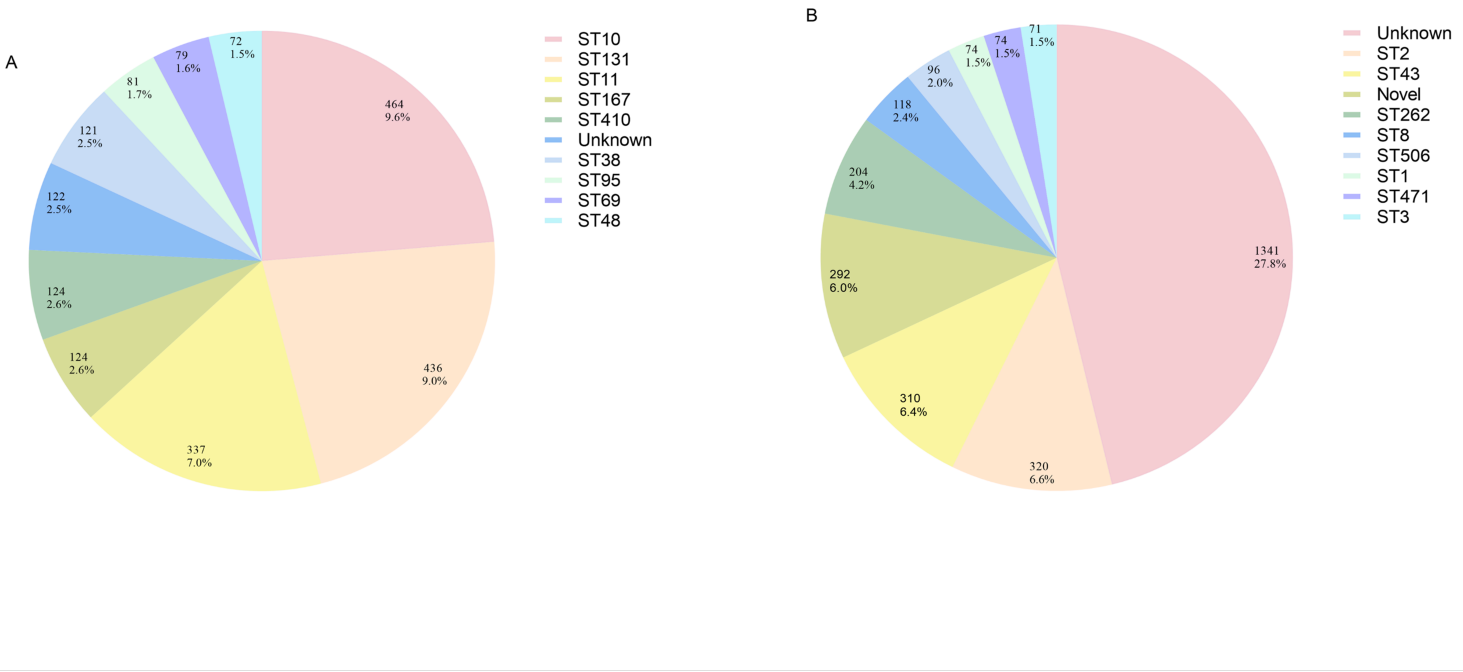


Figure 3

The proportion of the top 10 sequence types of *E. coil* strain by Atchman and Pasteur. (A) Sequence types classified by the Achtman scheme. (B) Sequence types classified by the Pasteur scheme. The pie charts display the prevalence of the most common STs, with the percentage and the number of isolates shown for each. STs labeled "Novel" represent new types, and "Unknown" indicates non-typeable isolates.

Supplementary Files

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

- Table1.docx
- Table2.docx
- Table3.docx
- Table4.docx
- TableS1.docx
- TableS2.docx
- TableS3.docx
- TableS4.docx
- TableS5.docx