

Article

Not peer-reviewed version

Diversity of CRISPR-Cas Systems Identified in Urological *Escherichia coli* Strains

Pavel V. Slukin , [Mikhail V. Fursov](#) , Daniil V. Volkov , Angelika A. Sizova , [Konstantin V. Detushev](#) , [Ivan A. Dyatlov](#) , [Nadezhda K. Fursova](#) *

Posted Date: 3 November 2025

doi: 10.20944/preprints202510.2396.v1

Keywords: CRISPR-Cas systems; uropathogenic *Escherichia coli* (UPEC); spacers sequences; type I-E; type I-F; virulence genes; antimicrobial resistance genes; whole-genome sequencing; genetic diversity



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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

Article

Diversity of CRISPR-Cas Systems Identified in Urological *Escherichia coli* Strains

Pavel V. Slukin, Mikhail V. Fursov, Daniil V. Volkov, Angelika A. Sizova, #
Konstantin V. Detushev, Ivan A. Dyatlov and Nadezhda K. Fursova *

State Research Center for Applied Microbiology and Biotechnology, 24 Territory "Kvartal A", 142279
Obolensk, Moscow region, Russia

* Correspondence: n-fursova@yandex.ru

Abstract

The type I-E and I-F CRISPR-Cas systems were identified in 237 *E. coli* strains isolated from patients with urinary tract infections (UTIs) between 2004 and 2019. The strains were classified into nine distinct groups (I-IX) based on the presence or absence of *cas* genes and repeat regions (RRs). Within the type I-E systems, two sequence variants were identified, distinguished by polymorphisms in the *casB*, *cas3*, *cas7*, *cas5*, and *cas6* genes. The direct repeats (DRs) also differed, with I-E-associated RRs ranging from 26–32 bp and I-F-associated RRs being a consistent 28 bp. We identified 762 unique spacers (29–35 bp in length) across the strain collection. The number of spacers per strain varied from 1 to 47, and potential DNA targets were determined for 83 spacers, targeting 56 bacteriophage genomes, 19 plasmids, and 8 *cas* genes of the I-F type. Multilocus sequence typing (MLST) revealed 68 sequence types and 24 clonal complexes (CCs), with ST131, CC10, CC69, CC405, CC14, CC38, CC73, and CC648 being the most prevalent. Significant correlations were observed between specific phylogroups/CCs, the type of CRISPR-Cas system present, and distinct profiles of virulence and antibiotic resistance genes.

Keywords: CRISPR-Cas systems; Uropathogenic *Escherichia coli* (UPEC); spacers sequences; # type I-E; type I-F; virulence genes; antimicrobial resistance genes; Whole-genome sequencing; genetic diversity

1. Introduction

Uropathogenic *E. coli* (UPEC) is the primary causative agent of urinary tract infections (UTIs), responsible for up to 90% of community-acquired and 50% of hospital-acquired cases [1]. The pathogenicity of UPEC is driven by a diverse arsenal of virulence genes involved in adherence, invasion, and iron acquisition. Furthermore, whole-genome sequencing (WGS) data confirmed the presence of a wide variety of resistance genes in *E. coli* genomes of this pathotype [2].

The CRISPR-Cas system, comprising Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins, which provides bacteria with an adaptive immune defense against foreign DNA, was first identified in *Escherichia coli* and studied in detail in the *E. coli* strain K-12. This foundational work established the functions of individual Cas proteins, elucidated the molecular mechanisms governing system assembly and activity, and identified key regulatory elements, including associated promoters and repressors [3,4]. In the past, the system was reported to be present in 45–50% of bacteria [5]. To date, additional functions of CRISPR-Cas systems have been identified, including the regulation of cryptic prophages [6,7].

In the *E. coli* K-12 strain, three CRISPR regions have been identified: *iap-cysH*, *clpA-infA*, and *ygcE-ygcF*. The first two regions contain *cas* genes, whereas the third comprises only RR sequences [8,9]. The *iap-cysH* and *clpA-infA* regions encode distinct sets of *cas* genes corresponding to the type I-E (*Ecoli* or CASS2) and the type I-F (*Ypest* or CASS3) systems, respectively [10,11]. Moreover, the type I-E system of *E. coli* K-12 has been shown to be similar to those of *Salmonella enterica* and *Klebsiella*

pneumoniae, whereas the type I-F system shares similarity with that of *Yersinia pestis* [11,12]. It has been suggested that the *iap-cysH* and *ygcE-ygcF* regions are more associated with protection against phage DNA, and the *clpA-infA* region with protection against plasmid DNA, but another study found the opposite relationship [13,14].

The nomenclature of CRISPR-Cas systems, particularly that of Cas proteins, has not been standardized. For example, the CasA protein is also referred to as Cse1 or Cas8e, and CasB as Cse2. Furthermore, there is no consistent nomenclature for RRs [9]. The type I-E CRISPR-Cas systems comprise eight Cas proteins: Cas3, CasA, CasB, Cas7, Cas5, Cas6, Cas1, and Cas2. Cas1 and Cas2 are involved in the processes of spacer acquisition and adaptation, whereas CasA, CasB, Cas7, Cas5, and Cas6 mediate CRISPR RNA processing and the assembly of the Cascade effector complex, which recognizes foreign nucleic acids. Finally, Cas3 cleaves the target DNA through its nuclease activity [15].

CRISPR arrays of CRISPR-Cas systems consist of direct repeats (DRs) and intervening spacers homologous to fragments of foreign DNA. Together, these elements form a repeat region (RR) [16,17]. Analysis of spacer targets across a large number of *E. coli* strains revealed that at least 59% of spacers are homologous to bacteriophage DNA [9]. Notably, spacers within *E. coli* CRISPR arrays frequently target DNA sequences derived from bacteriophage P7 and epidemiologically relevant plasmids such as pO111 and pO157 [9,18].

Although recent research on CRISPR-Cas systems has primarily focused on their applications in laboratory, medical, and industrial contexts, there is growing interest in elucidating their roles within natural populations of both pathogenic and nonpathogenic *E. coli*. For instance, associations between CRISPR-Cas systems and antimicrobial resistance (AMR) have been demonstrated in avian pathogenic *Escherichia coli*. Moreover, correlations between CRISPR-Cas systems, phylogroup classification, and sequence type have been observed in *E. coli* strains isolated from the birth canal of healthy women during the postpartum period. In addition, an association between CRISPR-Cas systems and virulence genes has been reported for UPEC [14,19–21].

This study focuses on the analysis of CRISPR-Cas systems in *E. coli* strains isolated from the urine of patients presenting with symptoms of urinary tract infection (UTI), with particular emphasis on their correlations with virulence factors, antibiotic resistance, and genetic lineage affiliation.

2. Materials and Methods

2.1. Strains and Growth Conditions

A total of 237 *E. coli* strains were obtained from the State Collection of Pathogenic Microorganisms "SCPM-Obolensk" and were isolated between 2004 and 2019 in the Central Region of the Russian Federation (Table S1). Bacteria were cultured on LB agar medium (Thermo Fisher Scientific, Waltham, MA, USA).

2.2. DNA Isolation and PCR Amplification

Genomic DNA for PCR was extracted by alkaline lysis. PCR amplification was carried out on a Mini Amp Plus instrument (Applied Biosystems Inc., Woburn, MA, USA) with the DreamTaq Green PCR Master Mix kit (Thermo Fisher Scientific, Waltham, MA, USA). Detection of specific genes and assignment to phylogenetic groups were performed as described by Clermont et al., 2015 [22].

2.3. Whole-Genome Sequencing

WGS was carried out on the Illumina MiSeq platform using the Nextera DNA Library Preparation Kit and MiSeq Reagent Kits v3 (Illumina, San Diego, CA, USA) following the manufacturer's instructions. The resulting single-end reads were assembled into contigs using SPAdes version 3.9.0 software (<https://cab.spbu.ru/software/spades/>, accessed 09 October 2025). The whole-genome sequences of 55 strains have been deposited in the GenBank database under BioProject accession number PRJNA269675 (Table S1).

2.4. Genomic Analysis

WGS were analyzed using the Center for Genomic Epidemiology web services MLST 2.0 and ResFinder 4.1 [23–25] (<http://www.genomicepidemiology.org/>, accessed 09 October 2025). Virulence genes of the UPEC pathotype strains were identified using the BLAST web service (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>, accessed 09 October 2025). Reference sequences of *E. coli* virulence genes were selected from four functional groups: (i) adhesins: *fimH* (GenBank AJ225176: 1340–2242), *yfcV* (CP015085.1: 1243313–1250325), *papGII* (AY212279.1), *afaA* (X76688.1: 3471–3776), *papGIII* (AY212281.1), *sfaS* (X16664: 22118–22609), and *focG* (DQ301498: 6744–7247); (ii) toxins: *hlyA* (M10133: 1320–4391), *usp* (AB027193), *cnf1* (X70670), and *vat* (NZ_JSGK01000039: 85437–89567); (iii) siderophores: *fyuA* (AM236324: 43579–45600), *chuA* (AP017620.1: 3851222–3853204), *iutA* (AY553855: 5766–7964), and *iroN* (X16664: 25805–27982); and (iv) protectins: *traT* (AY684127.1), *ompT* (KP657558.1), and *kpsMT* (AP024112.1: 870921–872368).

The primary identification of CRISPR-Cas systems was carried out using the CRISPR-Cas++ web service (<https://crisprcas.i2bc.paris-saclay.fr/>, accessed 09 October 2025), while refinement analysis was performed using the BLAST web service (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>, accessed 09 October 2025). Complete genomes of *E. coli* strains K-12 substr. MG1655 (U00096.3), B-8552 (JAHWDZ000000000.1), and B-8431 (SERV000000000.1) were used as reference sequences.

Phylogenetic analysis of whole-genome sequences was performed using Snippy version 4.6.0 with default settings. The resulting phylogenetic tree was visualized using the iTOL web resource (<https://itol.embl.de/>, accessed 09 October 2025).

2.5. Statistical Methods

Statistical analysis was performed using GraphPad Prism 8.0.1. One-way ANOVA with post-hoc Tukey’s multiple comparisons test was used. Significance was assumed at $p < 0.05$.

3. Results

3.1. General Characteristics of the Strains

A total of 237 *E. coli* strains were obtained from the State Collection of Pathogenic Microorganisms "SCPM-Obolensk". These strains were isolated between 2004 and 2019 from ten medical hospitals across Russia. The isolates originated from the urine of patients with the following clinical diagnoses: urinary tract infection of unknown localization ($n = 192$), chronic cystitis ($n = 25$), asymptomatic bacteriuria ($n = 6$), pyelonephritis ($n = 4$), chronic pyelonephritis ($n = 4$), cystitis ($n = 2$), urolithiasis ($n = 2$), gestational pyelonephritis ($n = 1$), and overactive bladder ($n = 1$).

The strains were attributed to seven phylogenetic groups of *E. coli*: B2 (54%), A (16%), D (15%), B1 (9%), F (2%), E (2%), and C (1%). Among 231 strains, 70 sequence types (STs) were identified according to the Warwick University scheme, belonging to 24 clonal complexes (CCs). CC131 was predominant ($n = 85$, 35%), while other common CCs included CC10 (ST10, ST167, ST617; 8%), CC69 (ST69; 5%), CC405 (ST405; 4%), CC14 (ST14, ST1193, ST1858; 3%), CC38 (ST38, ST315; 3%), CC73 (ST73, ST1154; 3%), and CC648 (ST648, ST3177; 3%). Sixteen rarer clonal complexes (CC23, CC101, CC155, CC59, CC95, CC350, CC12, CC46, CC86, CC156, CC165, CC168, CC349, CC354, CC394, and CC469) were each represented by 1–3 strains (Table S1).

Genetic determinants of antibiotic resistance were identified in 203 *E. coli* strains, with resistance observed to beta-lactams (80%), fluoroquinolones (70%), aminoglycosides (63%), sulfonamides (63%), tetracyclines (57%), phenicols (42%), macrolides (3%), polymyxins (2%), fosfomycins (1%), and ansamycins (1%). Virulence genes associated with the uropathogenic *E. coli* (UPEC) pathotype were detected in the strain genomes, including *fimH* (95%), *fyuA* (76%), *chuA* (68%), *traT* (68%), *iutA* (66%), *ompT* (64%), *yfcV* (55%), *usp* (53%), *kpsMTII* (46%), *hlyA* (25%), *iroN* (22%), *cnf1* (21%), *papGII* (17%), *vat* (14%), *afa* (12%), *papGIII* (10%), *sfaS* (7%), *focG* (4%), and *kpsMTIII* (3%) (Table S1).

3.2. Identification and Characterization of CRISPR-Cas Systems

Whole-genome sequences of 237 *E. coli* strains were analyzed for the presence of CRISPR-Cas system components. In total, 127 strains carried CRISPR-Cas systems, including 110 assigned to the type I-E and 17 to the type I-F. The remaining 110 strains lacked CRISPR-Cas systems (i.e., did not possess genes encoding Cas proteins), although 108 of these contained RRs (Figure 1).

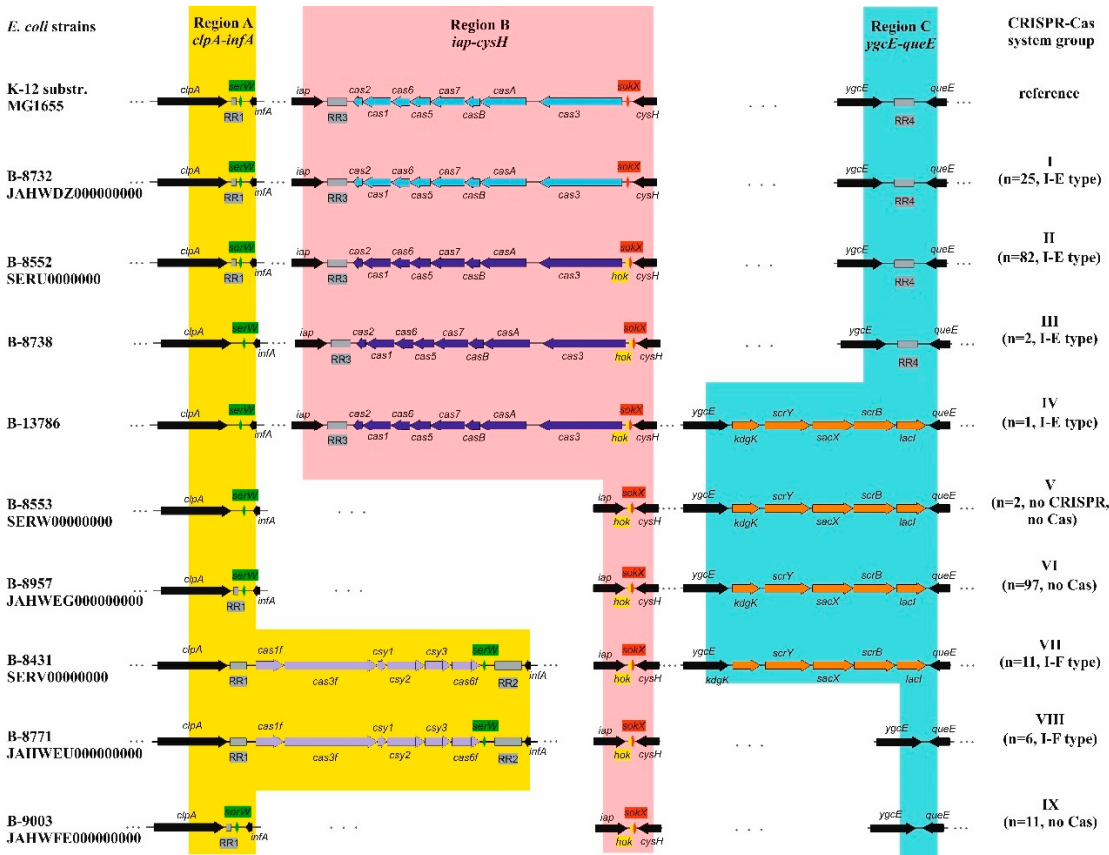


Figure 1. Genomic organization of CRISPR-Cas system components in uropathogenic *E. coli* strains compared to the reference strain K-12 substr. MG1655. Shaded areas represent distinct genetic regions: yellow (Region A), pink (Region B), and turquoise (Region C). Gray blocks denote RRs. Genes are indicated by arrows and color-coded as follows: green, *serW* (serine tRNA); red, *hok/sokX* (Hok toxin/SokX antitoxin module); light blue, type I-E Cas proteins (K-12 variant); blue, type I-E Cas proteins (B-8552 variant); purple, type I-F Cas proteins; orange, carbohydrate metabolism enzymes; black, genes unrelated to the CRISPR-Cas system.

Comparative analysis of whole-genome sequences identified three CRISPR-Cas-associated regions in the genomes of uropathogenic *E. coli* strains, which were analogous to those in the reference strain *E. coli* K-12 substr. MG1655. Region A is located between the *clpA* gene, a component of the Clp chaperone-protease operon, and the *infA* gene encoding translation initiation factor IF-1. Region B was found between the *iap* gene encoding alkaline phosphatase isoenzyme aminopeptidase, and the *cysH* gene for phosphoadenosine-phosphosulfate reductase. Region C was located between the sucrose kinase gene *ygcE* and *queE*, which encodes 7-carboxy-7-deazaguanine synthase. (Figure 1).

Region A contains a single RR1 in a large proportion (48%) of the studied strains (Groups I, II, VI, and IX). In 7% of strains, this region contains two distinct RRs, designated RR1 and RR2, which are separated by the type I-F *cas* gene cluster (Groups VII and VIII). A small subset of strains (2%) lacks any CRISPR-Cas structures in Region A (Groups III, IV, and V). Furthermore, the *serW* tRNA gene was identified between the *cas* gene cluster and RR2 in Groups VII and VIII. This gene is also

present in Region A of all other strains, including those lacking other CRISPR-Cas components (Figure 1).

Region B contains the RR3 locus and a set of the type I-E *cas* genes in 46% of strains (Groups I-IV). In the remaining strains (Groups V-IX), this region lacks CRISPR-Cas components. Additionally, Region B harbors genes of the Hok toxin-antitoxin system. In 90% of strains, this system is represented by both the *hok* toxin gene and the *sokX* antitoxin gene located adjacent to *cysH*, while in the remaining strains, only the *sokX* gene is present, similar to the reference *E. coli* K-12 substr. MG1655 strain (Figure 1).

Region C includes the RR4 locus in 46% of strains (Groups I-III). In 47% of strains (Groups IV-VII), this region carries a set of carbohydrate metabolism genes, including *kdgK* (encoding aminoimidazole riboside kinase), *scrY* (carbohydrate porin), *sacX* (subunit IIBC of the sucrose transporter), *scrB* (sucrose-6-phosphate hydrolase), and *lacI* (a transcriptional regulator). In the remaining strains (Groups VIII-IX), Region C lacks any genes (Figure 1).

3.3. Types of CRISPR-Cas Systems

Two types of CRISPR-Cas systems were identified among the studied uropathogenic *E. coli* strains, namely the type I-E (n = 110) located in genomic Region B, and the type I-F (n = 17) identified in genomic Region A (Figure 1).

The gene order and composition of the type I-E CRISPR-Cas systems (Groups I-IV) were as follows: *cas3-casA-casB-cas7-cas5-cas6-cas1-cas2*. However, based on sequence divergence, two distinct variants were identified. The first variant (var. K-12), identical to the reference strain *E. coli* K-12 substr. MG1655, was found in 11% of strains (Group I). The second variant (var. B-8552), identified in Groups II-IV, was characterized by substantial differences in the *cas3*, *cas7*, *cas5*, and *cas6* genes. These genes shared no significant nucleotide similarity with the var. K-12, while their encoded proteins exhibited 29-31% amino acid identity (with coverage >50%). In contrast, the *casA*, *cas1*, and *cas2* genes and their corresponding protein products showed high sequence similarity to those in var. K-12. The *casB* gene and its protein product were generally unique to each variant set (Figures S1-S7, Table 1).

The order and identity of Cas genes in type I-F CRISPR-Cas systems were as follows: *cas1f-cas3f-csy1-csy2-csy3-cas6f* (n = 17). Although the strains carrying this type of CRISPR-Cas system are divided into two groups (Groups VII and VIII), the distinction between them pertains solely to the genetic content of Region C, while the structure and composition of the I-F CRISPR-Cas system itself are identical in both (Figure 1).

Table 1. Comparison of nucleotide sequences of *cas* genes and amino acid sequences of the corresponding Cas proteins between the var. B-8552 and var. K-12 of the type I-E CRISPR-Cas system in uropathogenic *E. coli* strains.

Gene	var. B-8552 / var. K-12			
	Gene length, bp	GC-content, %	Gene QC / PI, %	Protein QC / PI, %
<i>cas3</i>	2700 / 2667	50 / 45	0 / 0	87 / 29
<i>casA</i>	1563 / 1509	50 / 44	1 / 94	53 / 23
<i>casB</i>	537 / 483	52 / 46	0 / 0	0 / 0
<i>cas7</i>	1056 / 1092	51 / 44	0 / 0	82 / 31
<i>cas5</i>	747 / 675	55 / 48	0 / 0	66 / 32
<i>cas6</i>	651 / 600	55 / 45	0 / 0	99 / 29
<i>cas1</i>	924 / 918	52 / 51	90 / 73	98 / 84
<i>cas2</i>	294 / 285	48 / 46	99 / 75	90 / 86

Note: QC / PI, Query Coverage / Percent Identity.

3.4. Analysis of the Repeat Regions, RRs

3.4.1. Direct Repeats, DRs

Analysis revealed that the DRs of RR1 and RR2, located within genomic Region A, consisted of identical 28 bp sequences (*tttctaagctgcctgtacggcagtgaac*) in the majority of strains (n = 232). In two strains, an additional cytosine was present at the right end of the sequence (Table S2).

The DRs within RR3 and RR4 (located in genomic Regions B and C, respectively) are 26-32 nucleotides in length. These sequences share a conserved 24-nucleotide core (*tttatcccgctggcgcggggaac*), which is flanked by 2-4 variable nucleotides on either end. Furthermore, some strains possess point mutations within this core sequence (Table S2, Figure 2).

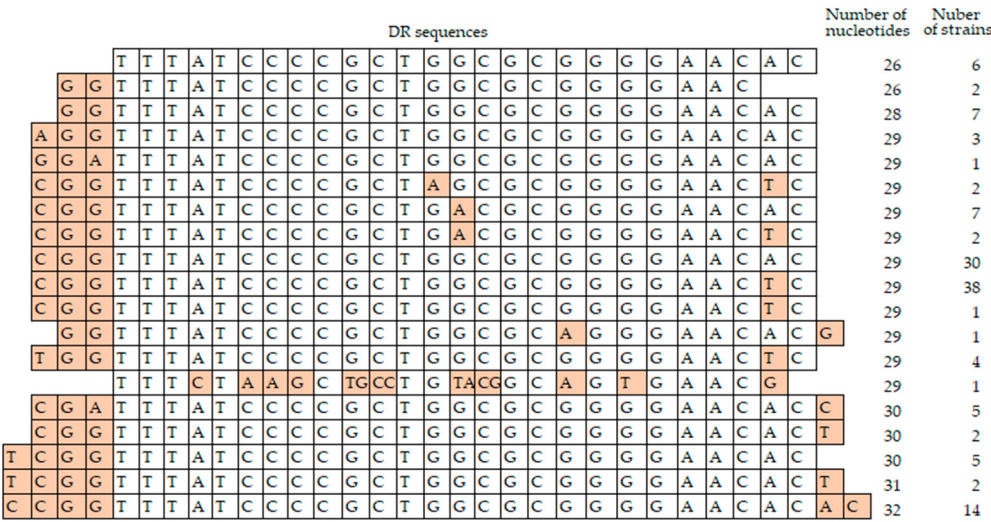


Figure 2. Comparison of DRs located in RR3 and RR4 regions. Variable nucleotides are highlighted in peach color.

3.4.2. Spacer Sets

A total of 762 unique spacers were identified across the studied urological *E. coli* strains. Spacer lengths ranged from 29 to 35 bp, with 32-bp spacers being the most prevalent (80%). Some spacers from different strains exhibited high sequence similarity, differing only by single nucleotide substitutions, and were therefore grouped into sequence clusters. The number of spacers per strain varied from 1 to 47, with two strains lacking spacers entirely. The most widely distributed spacers were sp876, sp665, and sp608, found in 41%, 23%, and 17% of strains, respectively. Notably, sp608 was present as two adjacent copies in six strains. The majority of spacers (42%) were unique, each identified in only a single strain (Tables S2 and S3).

Differences in the number of spacers among the RRs were observed: RR1 contained 1-14 spacers, RR2 2-33, RR3 2-30, and RR4 2-27. In total, 59 distinct spacers were identified in RR1, 50 in RR2, 355 in RR3, and 340 in RR4. The most prevalent spacers in RR1 were sp876 (n = 97), sp665 (n = 54), and sp608 (n = 46); in RR2, sp308 (n = 13), sp3 (n = 11), and sp11 (n = 11); in RR3, sp81 (n = 36), sp42 (n = 28), and sp250 (n = 17); and in RR4, sp151 (n = 26), sp398 (n = 24), sp49 (n = 23), and sp122 (n = 23). Notably, the majority of spacers (95%) were present in only one RR, 5% in two RRs, and only two spacers occurred in three RRs (Tables S2–S4).

BLAST analysis identified potential DNA targets for 83 spacers in the GenBank database. These targets were located in bacteriophage genomes (n = 56), plasmids (n = 19), and Cas protein coding sequences (n = 8). For instance, 28 spacers matched different regions of the *Escherichia* phage vB_EcoM-705R4 genome (GenBank ON470624.1), seven spacers matched the *Escherichia* phage PhiR41_1 genome (GenBank PV340561.1), four spacers matched the *Enterobacteria* phage P7 genome (GenBank NC_050152.1), and two spacers matched the *Lactacaseibacillus* phage R23.9 genome (GenBank OP869848.1) (Table S4, Figure 3).

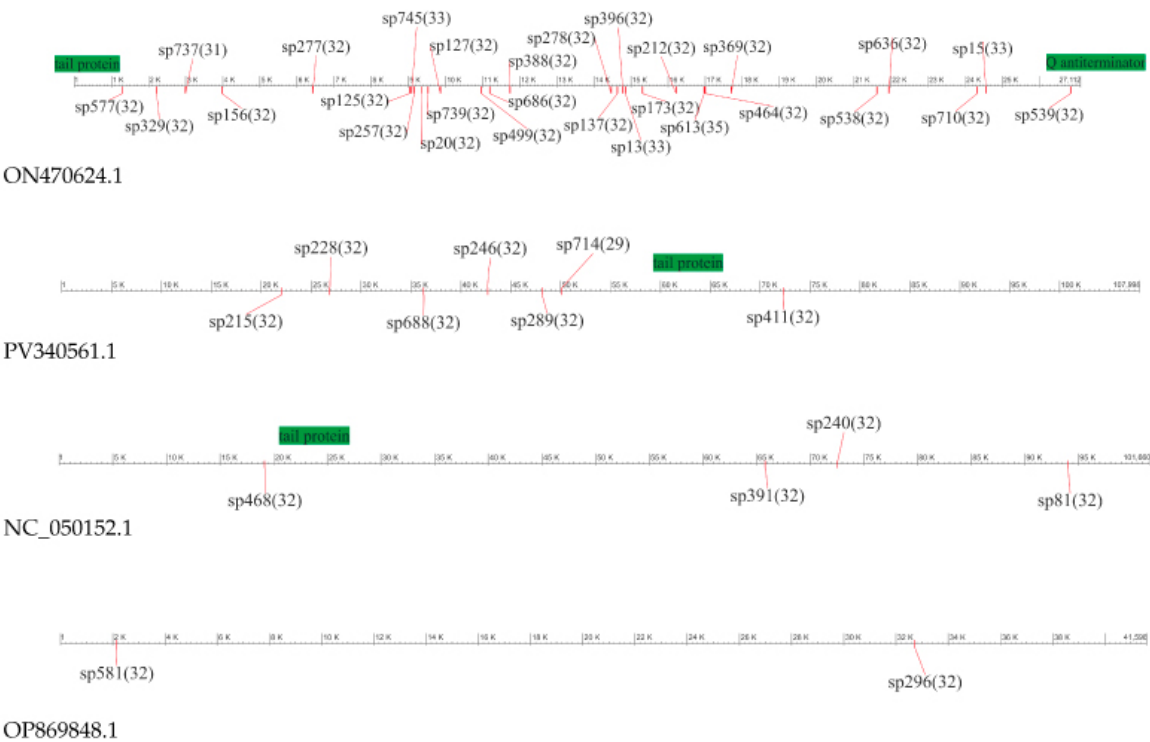


Figure 3. DNA targets for the spacers of the urological *E. coli* CRISPR-Cas systems in genomes of *Escherichia* phage vB_EcoM-705R4 (ON470624.1), *Escherichia* phage PhiR41_1 (PV340561.1), *Enterobacteria* phage P7 (NC_050152.1), and *Lacticaseibacillus* phage R23.9 (OP869848.1). Red stickers indicate the location of the target sites. The length of the target DNA sequence in bp is indicated in brackets.

Target DNAs for 19 spacers were identified on plasmids, including twelve spacers matched the plasmid pGF54-C (GenBank CP172162.1), seven spacers matched the plasmid p1-S1-IND-01-A (GenBank CP145658.1), six spacers matched the plasmid pEc2-51408 (GenBank CP104116.1), six spacers matched the plasmid pO157 (GenBank NZ_ABHM02000004.1), three spacers matched the plasmid p666 (GenBank FN649417.1), and two spacers matched the plasmid pO83_CORR (GenBank NC_017659.1). It is noteworthy that the target DNAs for seven spacers (sp40, sp228, sp246, sp270, sp289, sp347, and sp411) are localized in two different genomes, and for four spacers, on three plasmid replicons each (Table S4, Figure 4).

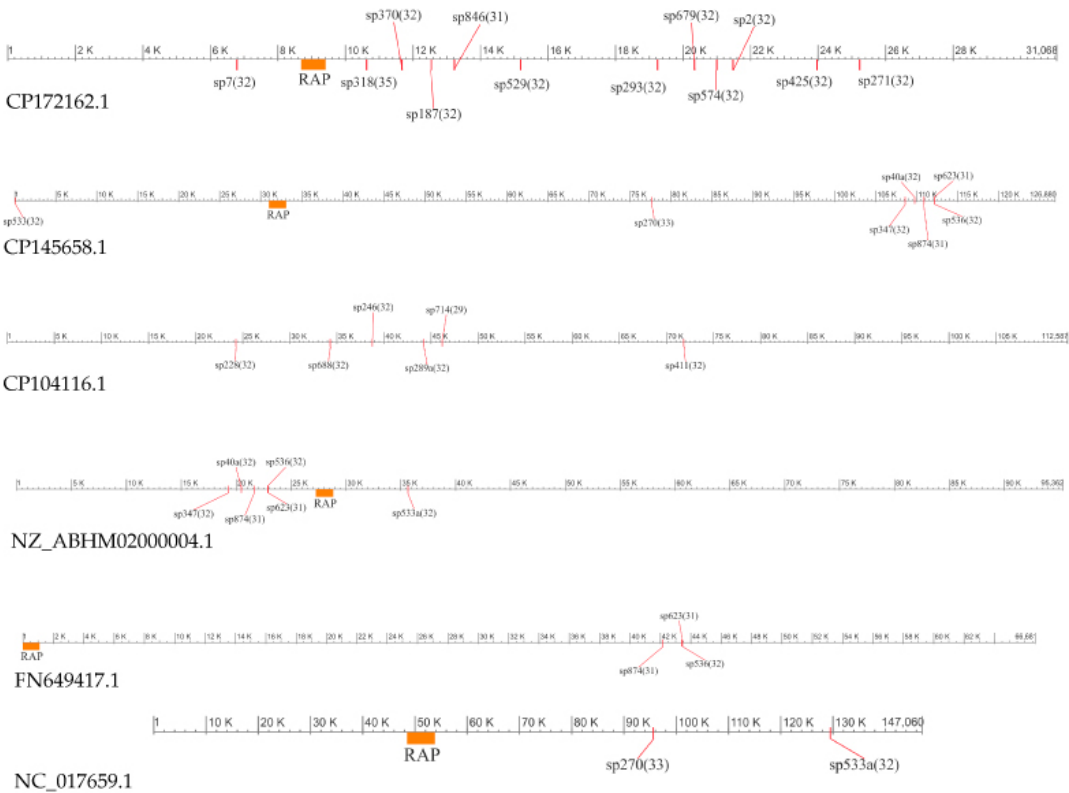


Figure 4. DNA targets for the spacers of the urological *E. coli* CRISPR-Cas systems on the plasmids pGF54-C (CP172162.1), p1-S1-IND-01-A (CP145658.1), pEc2-51408 (CP104116.1), pO157 (NZ_ABHM02000004.1), p666 (FN649417.1), and pO83_CORR (NC_017659.1). Red stickers indicate the location of the target sites. The length of the target DNA sequence in bp is indicated in brackets.

Furthermore, the target DNAs of eight spacers (sp28, sp487, sp491, sp608, sp665, sp673, sp875, and sp876) located in the RR1 of the studied *E. coli* strains of Groups I, II, VI, and IX (91% of strains) were identified in the sequences of the type I-F *cas* genes. Two spacers matched the *cas1f* gene, five spacers matched the *cas3f* gene, and one spacer matched the *csy3* gene (Table S4, Figure 5).

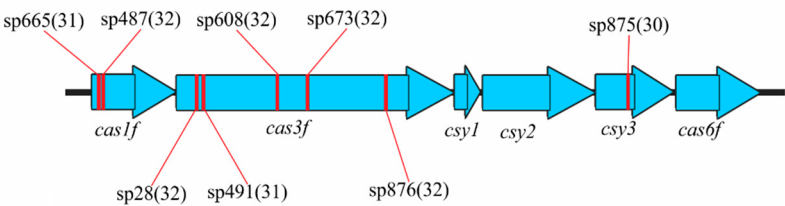


Figure 5. DNA targets for the spacers of the urological *E. coli* CRISPR-Cas systems in the *cas* genes of the type I-F CRISPR-Cas system. Red stickers indicate the location of the target sites. The length of the target DNA sequence in bp is indicated in brackets.

3.5. *E. coli* Phylogenetic Groups and CRISPR-Cas Systems Prevalence

Phylogenetic analysis of uropathogenic *E. coli* genomes revealed three major clusters. Strains from the CC131 formed a distinct, separate cluster, while the remaining strains were distributed between two heterogeneous groups comprising multiple clonal complexes and sequence types. A clear correlation emerged between phylogenetic clustering and the presence of specific CRISPR-Cas systems. For example, all strains belonging to the predominant clonal complex, CC131, lacked *cas* genes, although RR1 sequences were present in their genomes (Group VI). Conversely, every strain within the large cluster at the bottom of the phylogenetic tree (Figure 6) possessed a type I-E CRISPR-

Cas system (Groups I-IV). Similarly, all genomes encoding type I-F systems (Groups VII, VIII) were exclusively concentrated within a separate, distinct cluster (Figure 6).

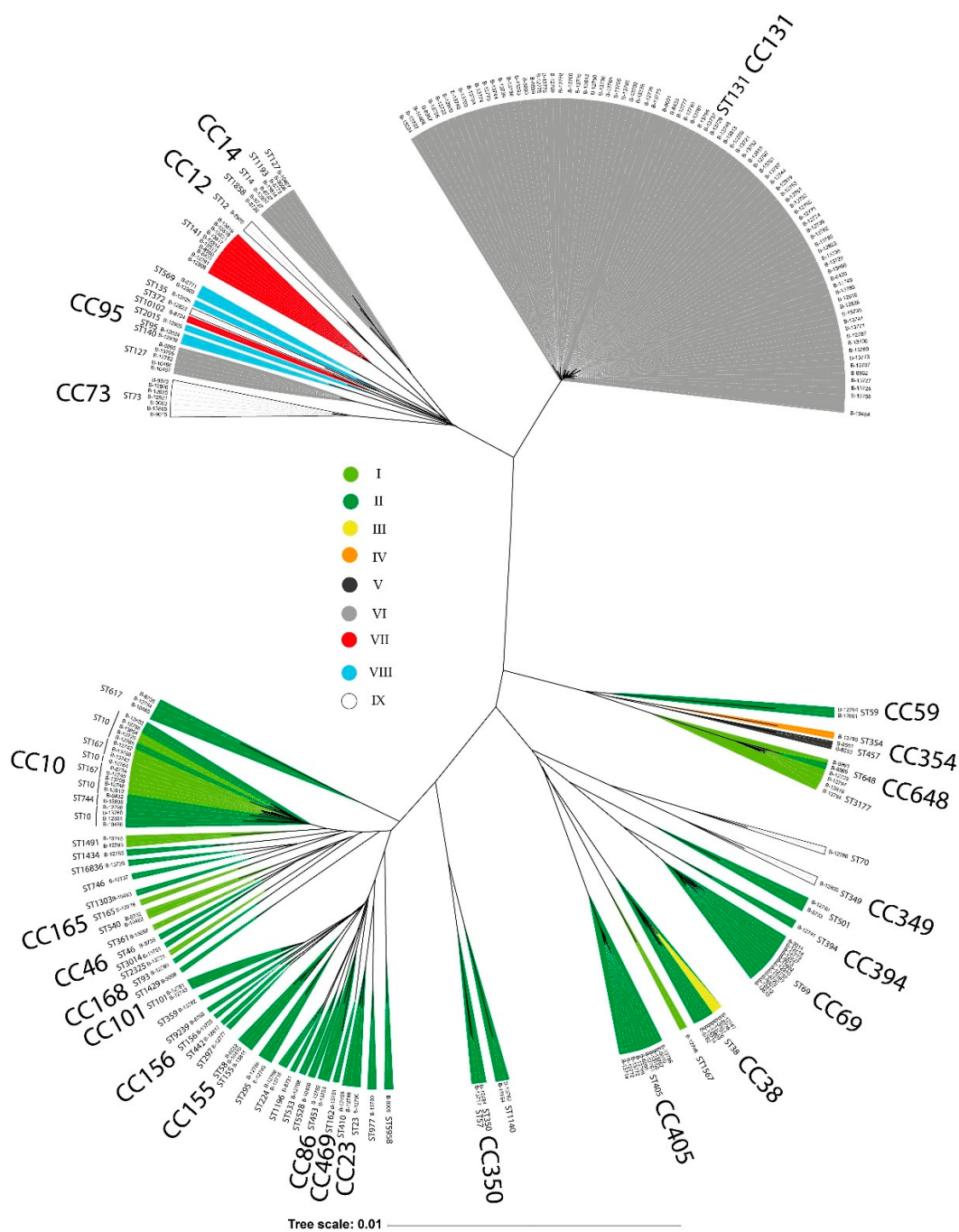


Figure 6. Phylogenetic tree constructed based on whole-genome sequences of urological *E. coli* strains using Snippy version 4.6.0 with default settings and visualized using the iTOL web resource.

3.6. Correlation of CRISPR-Cas System Type with Virulence Genes and AMR Genes

This analysis was performed on three strain cohorts: those carrying the type I-E CRISPR-Cas system, those harboring the type I-F system, and those lacking these systems. Strains in the I-E cohort carried significantly fewer virulence determinants overall compared to the other two cohorts. This trend was most pronounced for specific genes, including those encoding the adhesins SfaS, FocG, and YfcV; the siderophores IutA, FyuA, and ChuA; the toxins Cnf1, HlyA, Usp, and Vat; and the protectins OmpT and KpsMTII (Figure 7a).

The data demonstrate that a high abundance of virulence genes in UPEC strains is positively correlated with either the presence of a type I-E CRISPR-Cas system or the complete absence of any CRISPR-Cas system (Figure 7b).

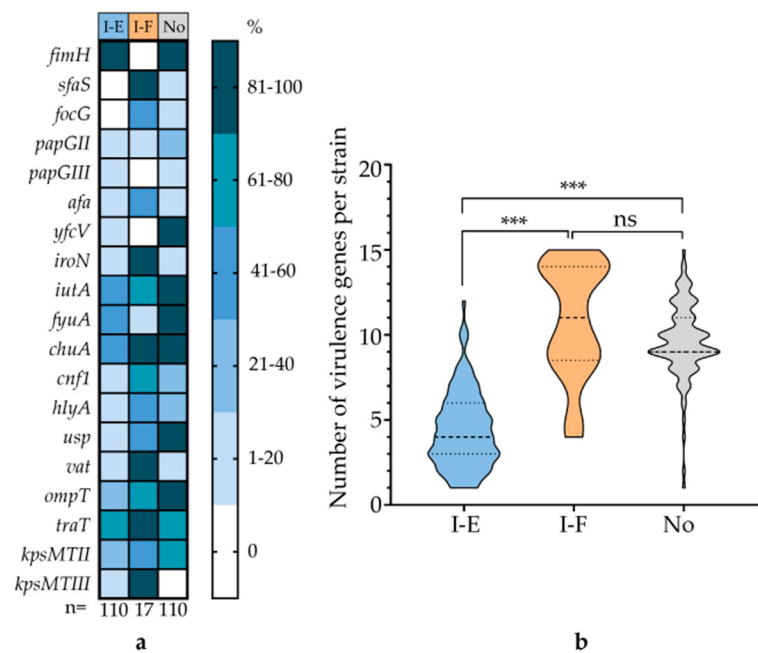


Figure 7. Correlation between CRISPR-Cas system types and virulence gene profiles in UPEC strains. **(a)** Heatmap depicting the prevalence of key UPEC virulence genes, categorized as adhesins (*fimH*, *sfaS*, *focG*, *papGII*, *papGIII*, *afaA*, *yfcV*), siderophores (*iroN*, *iutA*, *fyuA*, *chuA*), toxins (*cnfI*, *hlyA*, *usp*, *vat*), and protectins (*ompT*, *traT*, *kpsMTII*, *kpsMTIII*). **(b)** Statistical comparison of virulence gene content between strains carrying the type I-E systems, type I-F systems, or lacking a system (absence). Significance codes: ***, $p \leq 0.001$; ns, not significant ($p > 0.05$).

Analysis of AMR genetic determinants revealed that strains in the type I-F cohort carried significantly fewer AMR genes compared to the other two cohorts. These differences were most pronounced for genes conferring resistance to beta-lactams, aminoglycosides, sulfonamides, tetracyclines, fluoroquinolones, macrolides, and polymyxins (Figure 8a). Overall, these data demonstrate a significant positive correlation between AMR gene prevalence in UPEC strains and either the presence of type I-E CRISPR-Cas systems or the complete absence of CRISPR-Cas systems (Figure 8b).

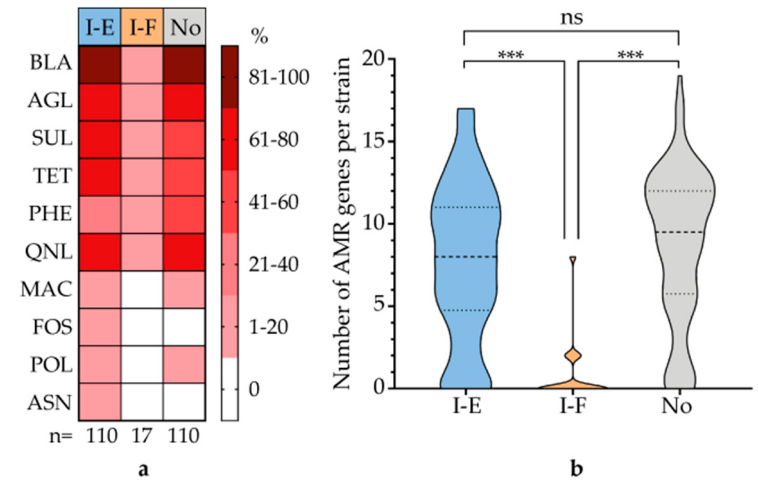


Figure 8. Correlation between CRISPR-Cas system types and AMR gene profiles in UPEC strains. **(a)** Heatmap depicting the prevalence of AMR genes associated with ten functional groups of antimicrobial agents: beta-lactams (BLA), aminoglycosides (AGL), sulfonamides (SUL), tetracyclines (TET), phenicols (PHE), fluoroquinolones (QNL), macrolides (MAC), fosfomycin (FOS), polymyxins (POL), and ansamycins (ANS). **(b)** Statistical comparison of AMR genes content between strains carrying the type I-E and type I-F CRISPR-Cas systems. Significance codes: ***, $p \leq 0.001$; ns, not significant ($p > 0.05$).

4. Discussion

In this study, we analyzed the whole-genome sequences of 237 uropathogenic *E. coli* (UPEC) strains collected from patients at ten medical hospitals across the Russian Federation between 2004 and 2019. Our analysis assigned the strains to seven phylogenetic groups (A, B1, B2, C, D, E, and F), 68 sequence types, and 24 clonal complexes (CCs). The most prevalent lineages included CC131, CC10, CC69, CC405, CC14, CC38, CC73, and CC648. The strains harbored a diverse repertoire of virulence genes encoding adhesins (*fimH*, *sfaS*, *focG*, *papGII*, *papGIII*, *afaA*, *yfcV*), siderophores (*iroN*, *iutA*, *fyuA*, *chuA*), toxins (*cnf1*, *hlyA*, *usp*, *vat*), and protectins (*ompT*, *traT*, *kpsMTII*, *kpsMTIII*). Antimicrobial resistance (AMR) profiling revealed that over 50% of strains carried genes conferring resistance to beta-lactams, fluoroquinolones, aminoglycosides, sulfonamides, tetracyclines, and phenicols, while fewer than 5% carried resistance determinants for macrolides, polymyxins, fosfomycins, and ansamycins. These genomic characteristics align with global epidemiological patterns of UPEC dissemination, virulence gene representation, and antibiotic resistance prevalence [2,26–28].

A significant characteristic of our collection was the presence of CRISPR/Cas systems in 53% of the strains. This prevalence is substantially higher than the 18% reported in a previous study from the USA [2]. The type I-E system was the most prevalent, identified in 46% of strains, while the type I-F system was less common, present in 7%. Phylogenetic analysis revealed that strains carrying type I-E systems were distributed across groups A, B1, B2, C, D, and F. In contrast, strains with type I-F systems or those lacking CRISPR-Cas systems predominantly belonged to phylogroup B2, consistent with published data [14].

The genomic architecture of CRISPR-Cas systems in the studied strains similar to that of the reference strain *E. coli* K-12 substr. MG1655, with components distributed across three designated regions (A, B, and C) [29]. Based on the presence or absence of *cas* genes, the composition of repeat regions (RR1-RR4), and additional genetic variations within these regions, the strains were categorized into nine distinct groups. These include strains carrying type I-E systems (Groups I-IV), type I-F systems (Groups VII-VIII), strains lacking *cas* genes but containing RRs (Groups VI, IX), and strains deficient in both *cas* genes and RRs (Group V) (Figure 1).

A key finding of our study is the identification of previously uncharacterized sequence divergence in the *cas3*, *cas7*, *cas5*, and *cas6* genes within type I-E CRISPR-Cas systems. The first variant (var. K-12), identical to the system in the reference strain *E. coli* K-12 substr. MG1655, was identified in 11% of strains (Group I). In contrast, a second, distinct variant (designated var. B-8552) was more prevalent, found in 35% of strains (Groups II-IV). Presumably, the *cas* genes of *E. coli* ST69 strains, which formed a separate clade among the *cas* genes of type I-E, described in a recent study, could be classified as having different variant of *cas* gene sequences [14].

Moreover, we identified incomplete *cas* operons (containing three, five, or six genes) in seven strains, consistent with previous reports [14]. These truncated operons consistently included *cas1*, *cas2*, and *cas3* genes, suggesting potentially diminished CRISPR-Cas activity. The occurrence of such deletions likely results from recombination events during bacterial evolution [2].

Further supporting the horizontal transfer of *cas* genes, we observed that closely related strains harbored different *cas* operon variants. Among ST10 strains, seven carried the var. K-12 system while six possessed the var. B-8552 variant. Similarly, ST648 strains included four strains with var. K-12 and one strain with var. B-8552 (Table S1).

Additional evidence for horizontal transfer comes from the identification of CRISPR-Cas components on plasmids in public databases, including: an unnamed 69,879 bp plasmid (CP089250.1) from an *E. coli* strain collected in South Korea, 2021; plasmid pIOMTU792 (LC542972.1) from a Nepalese human isolate, collected in 2020; and IncA/C2 plasmid p24C171-1 (LC501671.1) from a Japanese broiler isolate, collected in 2012.

Our analysis identified two distinct groups of direct repeats (DRs). The first group consisted of 28-bp DRs with the core sequence ttctaagctgcctgtacggcagtgaa located in genomic Region A (RR1 and RR2). The second group comprised 24-bp DRs with the core sequence ttatccccgctggcgcggggaac found in Regions B and C (RR3 and RR4). This observed polymorphism in CRISPR loci, particularly in DR sequences, has been previously documented in enterobacterial genomes [30].

We identified 762 unique spacers across the uropathogenic *E. coli* strain collection. Spacer counts per strain ranged from 1 to 47, with lengths varying between 29-35 bp. BLAST analysis revealed potential DNA targets for 83 spacers, including 56 bacteriophage genomes, 19 plasmids, and 8 *cas* genes. A significant number of spacers exhibited no significant similarity with sequences submitted to the GenBank database was pointed in other studies [2]. Interestingly, in *E. coli*, there appears to be a correlation of the distribution of the spacer number with pathogenic traits [31].

Our analysis confirms significant correlations between CRISPR-Cas system types and specific pathogenic traits. Strains carrying type I-F systems exhibited higher virulence gene content, while those with type I-E systems showed greater prevalence of antibiotic resistance determinants. The absence of a CRISPR-Cas system correlated with a elevated prevalence of both virulence and AMR genes. These findings align with existing literature [14,20] and support the potential of CRISPR-Cas analysis for elucidating mechanisms of bacterial pathogenicity and antimicrobial resistance [21].

5. Conclusions

In this study of UPEC strains collected from patients with urological infections in the Central region of Russia in 2004 - 2019, the prevalence of UPEC pandemic STs carrying a huge spectrum of virulence and AMR genes was shown. CRISPR/Cas systems were identified in 54% of the strains, including 46% of type I-E and 7% of type I-F CRISPR/Cas systems. Two variants of the type I-E CRISPR/Cas systems were identified based on the differences in *cas* genes sequences. A statistically significant correlation was found, on the one hand, between the I-E type CRISPR-Cas systems and antibiotic resistance genes, and, on the other hand, between the I-F type CRISPR-Cas systems and the virulence genes of uropathogenic *E. coli*. The obtained data expand our understanding not only of the prevalence of CRISPR-Cas systems among clinically significant UPEC strains but also indicate their role in the evolution of this pathogen.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, N.K.F. and P.V.S.; methodology, N.K.F.; software, D.V.V.; validation, D.V.V., A.A.S. and K.V.D.; formal analysis, M.V.F.; investigation, K.V.D. and P.V.S.; resources, I.A.D.; data curation, N.K.F.; writing—original draft preparation, P.V.S.; writing—review and editing, N.K.F. and M.V.F.; visualization, M.V.F.; supervision, I.A.D.; project administration, N.K.F.; funding acquisition, I.A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Higher Education of the Russian Federation, grant number 075-15-2025-525 of 30.05.2025. The APC was funded by Ministry of Science and Higher Education of the Russian Federation.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be

found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at <https://www.mdpi.com/ethics>.

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cas	CRISPR-associated proteins
RRs	Repeat Regions
DRs	Direct Repeats
UPEC	Uropathogenic <i>E. coli</i>
UTI	Urinary Tract Infection
WGS	Whole Genome Sequencing
ST	Sequence Type
CC	Clonal Complex

References

1. Toval, F.; Köhler, C.-D.; Vogel, U.; Wagenlehner, F.; Mellmann, A.; Fruth, A.; Schmidt, M.A.; Karch, H.; Bielaszewska, M.; Dobrindt, U. Characterization of *Escherichia coli* isolates from hospital inpatients or outpatients with urinary tract infection. *J. Clin. Microbiol.* **2014**, *52*, 407–418.
2. Gagaletsios, L.A.; Kikidou, E.; Galbenis, C.; Bitar, I.; Papagiannitsis, C.C. Exploring virulence characteristics of clinical *Escherichia coli* isolates from Greece. *Microorganisms*. **2025**, *13*(7), 1488.
3. Mitić, D.; Bolt, E.L.; Ivančić-Baće, I. CRISPR-Cas adaptation in *Escherichia coli*. *Biosci Rep.* **2023**, *43*(3), BSR20221198.
4. Mosterd, C.; Rousseau, G.M.; Moineau, S. A short overview of the CRISPR-Cas adaptation stage. *Can J Microbiol.* **2021**, *67*(1), 1-12.
5. Makarova, K.S.; Wolf, Y.I.; Alkhnbashi, O.S.; Costa, F.; Shah, S.A.; Saunders, S.J.; Barrangou, R.; Brouns, S.J.J.; Charpentier, E.; Haft, D.H.; et al. An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **2015**, *13*, 722–736.
6. Song, S.; Semenova, E.; Severinov, K.; Fernández-García, L.; Benedik, M.J.; Maeda, T.; Wood, T.K. CRISPR-Cas controls cryptic prophages. *Int J Mol Sci.* **2022**, *23*(24), 16195.
7. Dudley, E.G. The *E. coli* CRISPR-Cas conundrum: are they functional immune systems or genomic singularities? *EcoSal Plus.* **2025**, *9*, eesp00402020.
8. Díez-Villaseñor, C.; Almendros, C.; García-Martínez, J.; Mojica, F.J. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology (Reading)*. **2010**, *156*(5), 1351-1361.
9. Dion, M.B.; Shah, S.A.; Deng, L.; Thorsen, J.; Stokholm, J.; Krogfelt, K.A.; Schjørring, S.; Horvath, P.; Allard, A.; Nielsen, D.S.; Petit, M.A.; Moineau, S. *Escherichia coli* CRISPR arrays from early life fecal samples preferentially target prophages. *ISME J.* **2024**, *18*(1), wrae005.
10. Makarova, K.S.; Haft, D.H.; Barrangou, R.; Brouns, S.J.; Charpentier, E.; Horvath, P.; Moineau, S.; Mojica, F.J.; Wolf, Y.I.; Yakunin, A.F.; van der Oost, J.; Koonin, E.V. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol.* **2011**, *9*(6), 467-477.
11. Touchon, M.; Charpentier, S.; Clermont, O.; Rocha, E.P.; Denamur, E.; Branger, C. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J Bacteriol.* **2011**, *193*(10), 2460-7.

12. Iordache, D.; Baci, G.M.; Căpriță, O.; Farkas, A.; Lup, A.; Butiuc-Keul, A. Correlation between CRISPR loci diversity in three enterobacterial taxa. *Int J Mol Sci.* **2022**, *23*(21), 12766.
13. Touchon, M.; Rocha, E.P. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One.* **2010**, *5*(6), e11126.
14. Mikhaylova, Y.; Tyumentseva, M.; Karbyshev, K.; Tyumentsev, A.; Slavokhotova, A.; Smirnova, S.; Akinin, A.; Shelenkov, A.; Akimkin, V. Interrelation between pathoadaptability factors and Crispr-element patterns in the genomes of *Escherichia coli* isolates collected from healthy puerperant women in Ural region, Russia. *Pathogens.* **2024**, *13*(11), 997.
15. Yoshimi, K.; Takeshita, K.; Koder, N.; Shibumura, S.; Yamauchi, Y.; Omatsu, M.; Umeda, K.; Kunihiro, Y.; Yamamoto, M.; Mashimo, T. Dynamic mechanisms of CRISPR interference by *Escherichia coli* CRISPR-Cas3. *Nat Commun.* **2022**, *13*(1), 4917.
16. Sontheimer, E.J.; Barrangou, R. The bacterial origins of the CRISPR genome-editing revolution. *Hum Gene Ther.* **2015**, *26*(7), 413-424.
17. Tajkarimi, M.; Wexler, H.M. CRISPR-Cas systems in *Bacteroides fragilis*, an important pathobiont in the human gut microbiome. *Front Microbiol.* **2017**, *23*(8), 2234.
18. Sheludchenko, M.S.; Huygens, F.; Stratton, H.; Hargreaves, M. CRISPR diversity in *E. coli* isolates from Australian animals, humans and environmental waters. *PLoS One.* **2015**, *10*(5), e0124090.
19. Dong, H.; Cui, Y.; Zhang, D. CRISPR/Cas technologies and their applications in *Escherichia coli*. *Front Bioeng Biotechnol.* **2021**, *9*, 762676.
20. Kim, K.; Lee, Y.J. Relationship between CRISPR sequence type and antimicrobial resistance in avian pathogenic *Escherichia coli*. *Vet Microbiol.* **2022**, *266*, 109338.
21. Dziuba, A.; Dzierżak, S.; Sodo, A.; Wawszczak-Kasza, M.; Zegadło, K.; Białek, J.; Zych, N.; Kiebzak, W.; Matykiewicz, J.; Głuszek, S.; Adamus-Białek, W. Comparative study of virulence potential, phylogenetic origin, CRISPR-Cas regions and drug resistance of *Escherichia coli* isolates from urine and other clinical materials. *Front Microbiol.* **2023**, *14*, 1289683.
22. Clermont, O.; Gordon, D.; Denamur, E. Guide to the various phylogenetic classification schemes for *Escherichia coli* and the correspondence among schemes. *Microbiology.* **2015**, *161*, 980–988.
23. Bortolaia, V.; Kaas, R.S.; Ruppe, E.; Roberts, M.C.; Schwarz, S.; Cattoir, V.; Philippon, A.; Allesoe, R.L.; Rebelo, A.R.; Florensa, A.R.; et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother.* **2020**, *75*(12), 3491-3500.
24. Clausen, P.T.L.C.; Aarestrup, F.M.; Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics.* **2018**, *19*(1), 307.
25. Larsen, M.; Cosentino, S.; Rasmussen, S.; Rundsten, C.; Hasman, H.; Marvig, R.; Jelsbak, L.; Sicheritz-Ponten, T.; Ussery, D.; Aarestrup, F.; Lund, O. Multilocus sequence typing of total genome sequenced bacteria. *J Clin Microbiol.* **2012**, *50*(4), 1355-1361.
26. Joyce, S.; Belmont, C.; Scheffler, A.W.; Ravi, K.; Kim, H.; Rubin-Saika, N.; Elises, M.; Soto, A.; Mahesh, P. A.; Chambers, H.; Raphael, E. Trends in uropathogenic *Escherichia coli* genotype and antimicrobial resistance from 2019 to 2022 in a San Francisco Public Hospital Network. *Open Forum Infect Dis.* **2025**, *12*(9), ofaf579.
27. Gebremedhin, K.B.; Amogne, W.; Alemayehu, H.; Bopegamage, S.; Eguale, T. The role of uropathogenic *Escherichia coli* virulence factors in the development of urinary tract infection. *J Med Life.* **2025**, *18*(8), 701-709.
28. Nasrollahian, S.; Graham, J.P.; Halaji, M. A review of the mechanisms that confer antibiotic resistance in pathotypes of *E. coli*. *Front Cell Infect Microbiol.* **2024**, *14*, 1387497.
29. Almendros, C.; Mojica, F.J.; Díez-Villaseñor, C.; Guzmán, N.M.; García-Martínez, J. CRISPR-Cas functional module exchange in *Escherichia coli*. *mBio.* **2014**, *5*(1), e00767-13.
30. Iordache, D.; Baci, G.M.; Căpriță, O.; Farkas, A.; Lup, A.; Butiuc-Keul, A. Correlation between CRISPR loci diversity in three enterobacterial taxa. *Int J Mol Sci.* **2022**, *23*(21), 12766.
31. García-Gutiérrez, E.; Almendros, C.; Mojica, F.J.M.; Guzmán, N.M.; García-Martínez, J. CRISPR content correlates with the pathogenic potential of *Escherichia coli*. *PLoS ONE.* **2015**, *10*, e0131935.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim

responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.