Virulence and antimicrobial resistance genes are enriched in the plasmidome of clinical *Escherichia coli* isolates compared with wastewater isolates from western Kenya

Sifuna Anthony Wawire[a], Oleg N. Reva[b], Thomas J. O'Brien[c], Wendy Figueroa[c], Victor Dinda[d], William A. Shivoga[e], Martin Welch[c*]

Affiliations:

a. Department of Biochemistry, Masinde Muliro University of Science and Technology. P.O. Box 150, 50100 Kakamega, Kenya

b. Centre for Bioinformatics and Computational Biology; Dep. of Biochemistry, Genetics and Microbiology; University of Pretoria, Lynnwood Rd, Hillcrest, Pretoria 0002, South Africa

c. Department of Biochemistry, University of Cambridge; Hopkins Building, Tennis Court Road, Cambridge CB21QW, United Kingdom

d. Department of Medical Laboratory Science, Masinde Muliro University of Science and Technology. P.O. Box 150, 50100 Kakamega, Kenya

e. Department of Biological Sciences, Masinde Muliro University of Science and Technology. P.O. Box 150, 50100 Kakamega, Kenya

For correspondence:

*Martin Welch: Department of Biochemistry, University of Cambridge; Hopkins Building, Tennis Court Road, Cambridge CB21QW, United Kingdom. Email: mw240@cam.ac.uk

Declarations of interest:

None.

Highlights

- The physical location (plasmidome or chromosome) of genes involved in AMR and virulence is different in clinical and wastewater isolates.
- The chromosomal genes associated with virulence and antimicrobial resistance are often located on genomic islands.
- Maintenance of plasmid-borne functions contribute to the success of ST43.

Abstract

Many low-middle income countries in Africa have poorly-developed infectious disease monitoring systems. Here, we employed whole genome sequencing (WGS) to investigate the presence/absence of antimicrobial resistance (AMR) and virulence-associated (VA) genes in a collection of clinical and municipal wastewater *Escherichia coli* isolates from Kakamega, west Kenya. We were particularly interested to see whether, given the association between infection and water quality, the isolates from these geographically-linked environments might display similar genomic signatures. Phylogenetic analysis based on the core genes common to all of the isolates revealed two broad divisions, corresponding to the commensal/enterotoxigenic *E. coli* on the one hand, and uropathogenic *E. coli* on the other. Although the clinical and wastewater isolates each contained a very similar mean number of antibiotic resistance-encoding genes, the clinical isolates were enriched in genes required for in-host survival. Furthermore, and although the chromosomally encoded repertoire of these genes was similar in all sequenced isolates, the genetic composition of the plasmids from clinical and wastewater *E. coli* was more habitat-specific, with the clinical isolate plasmidome enriched in AMR and VA genes. Intriguingly, the plasmid-borne VA genes were often duplicates of genes already present on the chromosome, whereas the plasmid-borne AMR determinants were more specific. This reinforces the notion that plasmids are a primary means by which infection-related AMR and VA-associated genes are acquired and disseminated among these strains.

Keywords: *Escherichia coli*; genomics; antimicrobial resistance; virulence; plasmid; genomic islands.

1. Introduction

Intestinal pathogenic *Escherichia coli* are the leading bacterial cause of diarrheal infection in low-middle income countries (LMICs) (Jafari, et al., 2012). This situation is complicated by poorly-developed infectious disease monitoring programs, especially with regards to antimicrobial resistance (AMR) and pathogen surveillance (Vernet et al., 2014). In many LMICs, the lack of high-quality data frequently leads to inadequate treatment guidelines and poor infection management. One potentially transformative technology that could help improve pathogen surveillance programs in developing countries is whole genome sequencing (WGS) achieved through next generation sequencing (NGS) technologies. When coupled with epidemiological and environmental investigations, WGS can deliver ultimate resolution for detecting and analysing transmission routes and in tracing the source(s) of epidemics and outbreaks (Cao et al., 2017; Besser et al., 2018; Rantsiou et al., 2018). In addition, WGS offers the unrivalled opportunity to monitor the gene content of microbial virulence determinants in isolates, and to map the spread of antimicrobial drug resistance determinants (European Centre for Disease Prevention and Control, 2018). The technology is also particularly good at identifying mobile genetic elements such as plasmids, transposons and integrons, which are increasingly recognized as playing a key role in disseminating AMR and virulence determinants (Bezuidt et al., 2011).

In the current study, we employed WGS to monitor the genetic structure of *E. coli* recovered from patients visiting a referral health facility and from a nearby wastewater treatment plant in Kakamega, western Kenya. Previous studies have shown that the wastewater treatment sites in Kakamega are inefficient in controlling or removing pathogens from the water-supply system (Malaho et al., 2018). This inefficient wastewater treatment may therefore plausibly provide means by which multi-drug resistant and/or pathogenic *E. coli* are disseminated among the local populace. Phylogenic relationships between the *E. coli* isolates associated with nosocomial infections and wastewater treatment sites were inferred, and multilocus sequence typing of the isolates was performed. This revealed over-representation of one particular sequence type (ST 43) in the wastewater and clinical isolates. We also paid particular attention to the relative distribution of virulence-associated (VA) genes and antimicrobial resistance (AMR) genes on mobile genetic elements. Various VA and AMR genes were found in genomic islands (GIs) and plasmids of the sequenced strains. However, whereas VA and AMR genes were distributed among the GIs in both clinical and wastewater isolates, there was a marked enrichment of these genes in the plasmid DNA borne by the clinical isolates. This suggests that plasmid-mediated horizontal gene transfer may play a key role in defining the pathogenicity of these geographically linked *E. coli* isolates.

2. MATERIALS AND METHODS

2.1. Strains, Isolation, and Culture Conditions

The *E. coli* strains from human were recovered from patients being treated at the Kakamega County Teaching and Referral Hospital, in western Kenya. Briefly, midstream urine specimens were cultured on cystine–lactose–electrolyte-deficient (CLED) medium (Hi-Media, India). The wound sample was collected from an abdominal surgical wound. Wastewater isolates were recovered from the Masinde Muliro University of Science and Technology (MMUST) wastewater treatment plant during the period March – June 2016. The wastewater and wound-derived samples were cultured on MacConkey agar (Hi-Media, India).

All cultures were incubated overnight at 37^{o}C. Pure (single colony) isolates were confirmed as *E*. *coli* using API20E biochemical test strips (Biomerieux) following the manufacturer's instructions. The clinical isolates were further subjected to antibiotic susceptibility profiling using the Kirby Bauer disc diffusion method against the following panel of antibiotics: trimethoprim/sulfamethoxazole, amoxicillin/clavulanate, tetracycline, gentamicin, ceftazidime, cefuroxime, cefotaxime, ceftriaxone, meropenem, amikacin/cefepime, piperacillin/tazobactam, ampicillin, sulbactam, nitrofurantoin, imipenem, and ciprofloxacin. The environmental isolates were tested against trimethoprim/sulfamethoxazole, amoxicillin, amoxicillin/clavulanate, gentamicin, chloramphenicol, cephalexin, cefuroxime and ciprofloxacin. *Escherichia coli* ATCC 25922 was used as a reference for interpretation of the data based on the guidelines from the Clinical and Laboratory Standards Institute (CLSI), (2017). Cultures of the *E. coli* isolates were stored at −80^{o}C in trypticase soy broth supplemented with 15% v/v glycerol.

2.2. DNA sequencing

WGS was carried out by MicrobesNG (Birmingham, UK). Briefly, a single colony of each strain was picked and suspended in 100 µL of sterile 1 × phosphate-buffered saline (PBS) (Oxoid, UK). The suspension was spread thickly (using a sterile loop) onto a fresh LB-agar plate and incubated at 37^{o}C overnight. Dense colony growth was then scraped off and sent to MicrobesNG in supplied bar-coded bead tubes. Sequencing was carried out using an Illumina HiSeq 2500 platform, with 2 x 250 bp paired-end reads. The reads were trimmed using Trimmomatic v0.30 with a sliding window quality cut-off of Q15. Taxonomic classification of the sequences and assessment of sequence contamination was done using Kraken (Wood and Salzberg, 2014). The *de novo* assembly of contigs was done using SPAdes version 3.14.0 with default settings.

2.3. Annotation of contigs and prediction of chromosomal or plasmid affiliation

Automated annotation of the contigs was performed using Prokka v1.12. Antibiotic resistance genes were predicted using an internet-based algorithm, RGI (https://card.mcmaster.ca/analyze) implemented in the Web-service CARD (Alcock et al., 2020). Putative virulence-associated genes were identified *via* BLASTP alignment of all translated open reading frames (ORFs) against the sequences of known virulence associated proteins (Sarowska et al., 2019). Reference protein sequences were obtained from the NCBI GenBank database. To predict the putative plasmid affiliation of the assembled contigs, an internet-based program, mlplasmids - version 1.0.0 (https://sarredondo.shinyapps.io/mlplasmids/) was used (Arredondo-Alonso et al., 2018) with default parameters. All other contigs were considered as being chromosomal in origin. Genomic islands were identified using SeqWord Gene Island Sniffer (http://seqword.bi.up.ac.za/sniffer/index.html) (Bezuidt et al., 2009).

2.4. Multi-locus sequence typing

MLST was carried out using the interactive batch sequence query available at the Institute Pasteur *E. coli* MLST database (https://bigsdb.pasteur.fr/ecoli/) and using the automated interactive MLST CGE Server (https://cge.cbs.dtu.dk/services/MLST/) (Larsen et al., 2012). Concatenated chromosomal contigs in FASTA format were queried against the MLSTwithMissingData database to predict sequence types based on the sequences of eight diagnostic genes: *dinB*, *icdA*, *pabB*, *polB*, *trpA*, *trpB* and *uidA*, which are defined as *E. coli* MLST marker loci in the EnterBase database (Zhou et al., 2020). Sequence types (ST) unambiguously predicted by at least one of the servers were recorded.

2.5. Gene ortholog prediction

Clusters of orthologous genes (COGs) in sequenced genomes were predicted using the program OrthoFinder (Emms and Kelly, 2015) with default parameters. Sequences of every COG were aligned using the MUSCLE algorithm (Edgars 2004). Alignments were quality-controlled and ambiguous parts of the alignments were removed using the program Gblocks (Talavera and Castresana, 2007) with the default parameter settings. COG alignments were concatenated using BioPython scripts into a superstring alignment for further phylogenetic inferences.

2.6. Phylogenetic inferencing and clustering

Concatenated alignments of COGs were used to infer phylogenetic relations between the genome sequences. This was done using a Neighbor-Joining (NJ) algorithm with MEGA X and a bootstrap value of 100 (Kumar et al., 2018). For clustering of genomic islands (GIs), a distance matrix was built for all GIs, where the distance between two GIs, *i* and *j*, was calculated according to equation 1:

$$D_{ij} = 1 - \#shared\_genes \,/\, \min(\#GI_i\, genes, \#GI_j\, genes)$$

Eqn. 1. Where, *#shared_genes* is the number of orthologous genes shared by two GIs, and $\#GI_i\,genes$ and $\#GI_j\,genes$, respectively, are the total numbers of genes in the first and second GIs. A dendrogram of GI clusters was then generated based on the distance matrix using the program *neighbor.exe* (NJ algorithm) of the PHYLIP 3.69 package (http://evolution.genetics.washington.edu/phylip.html).

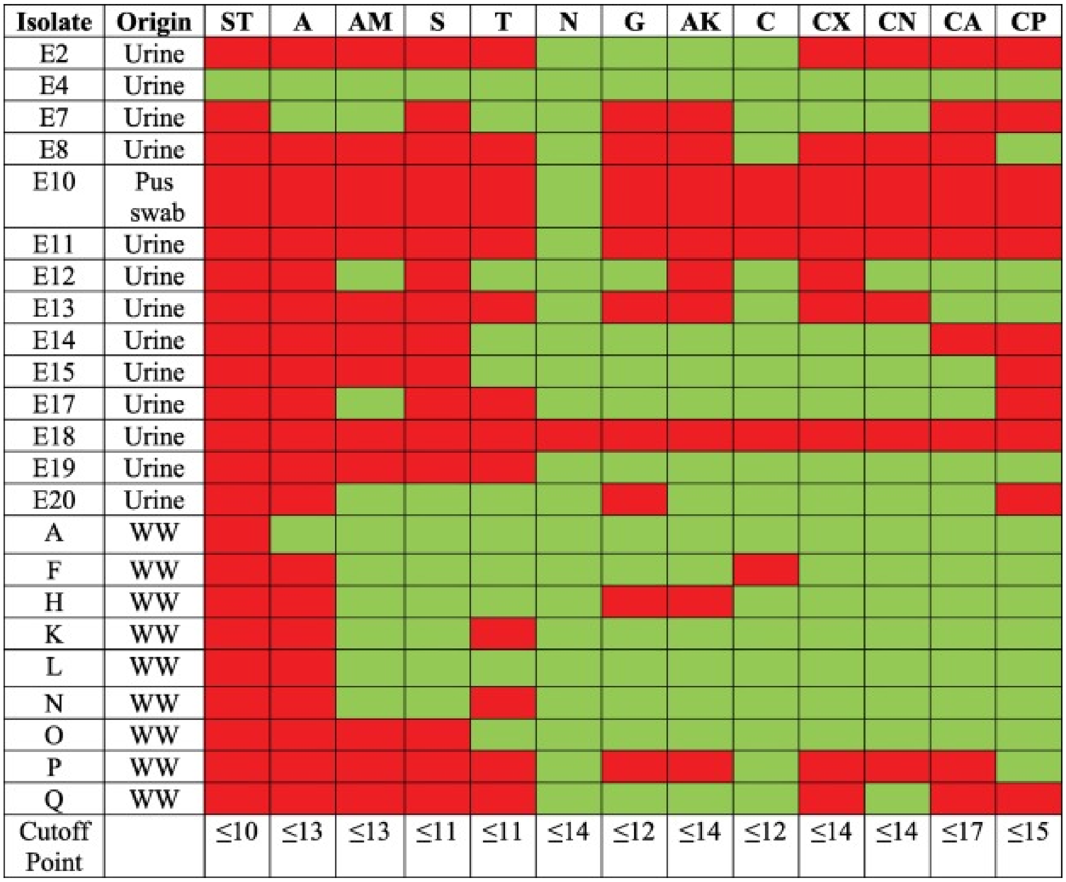
The phylogenetic clustering approach outlined above tends to artificially group long plasmids around the root of the dendrogram, as they have a higher chance of sharing multiple genes. To circumvent this, phylogenetic clustering of the plasmids was done using a binary parsimony algorithm. A table of COGs shared by at least two plasmids was generated, with the absence and presence of orthologous genes in each plasmid designated by 0 and 1, respectively. This binary table was formatted as an input file for the dollo parsimony algorithm (Huson and Steel, 2004) implemented in the program *dollop.exe* of the PHYLIP 3.69 package. Dendrograms were visualized using Dendroscope 3 (Huson and Scornavacca, 2012). It should be noted that in both aforementioned cases, the dendrograms should not (in *stricto sensu*) be considered as phylogenetic trees but as cladograms.

3. RESULTS

3.1. Antibiotic resistance of *E. coli* environmental and clinical isolates

In total, 23 strains of *E. coli* were recovered from a referral hospital and a nearby municipal wastewater processing plant in western Kenya. The source and pattern of antibiotic resistance in each of the strains are shown in Table 1. Nine isolates (denoted A, F, H, K, L, N, O, P, Q) were derived from the wastewater source (located ca. 1.5 km from the referral hospital) and 14 isolates were obtained from the clinic (denoted E2, E4, E7, E8, E10-E15, and E17-E20). Of the clinical isolates, 13 were obtained from patients with urinary tract infections, and one (E10) was from a wound. Almost all isolates from both the clinical and wastewater sources were resistant to trimethoprim-sulfamethoxazole and amoxicillin, and the clinical isolates showed generally greater resistance (compared with the wastewater isolates) to co-amoxiclav (amoxicillin/clavulanate), ampicillin-sublactam, tetracycline, gentamicin, amikacin, cefuroxime, cephalexin, ceftazidime, and ciprofloxacin. The majority of isolates from both environments remained resistant to nitrofurantoin and chloramphenicol.

Table 1. Antibiotic resistance profile and origin of the *E. coli* strains selected for WGS analysis in this study.



Isolate	Origin	ST	A	AM	S	T	N	G	AK	C	CX	CN	CA	CP
E2	Urine	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥
E4	Urine	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
E7	Urine	🟥	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟥	🟥
E8	Urine	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟩
E10	Pus swab	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
E11	Urine	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
E12	Urine	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟩	🟩	🟩
E13	Urine	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟥	🟥	🟩	🟩
E14	Urine	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥
E15	Urine	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
E17	Urine	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥
E18	Urine	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
E19	Urine	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
E20	Urine	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥
A	WW	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
F	WW	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩
H	WW	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟩	🟩
K	WW	🟥	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
L	WW	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
N	WW	🟥	🟥	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
O	WW	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
P	WW	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟩
Q	WW	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟩	🟥	🟥
Cutoff Point		≤10	≤13	≤13	≤11	≤11	≤14	≤12	≤14	≤12	≤14	≤14	≤17	≤15

Key: WW = wastewater; ST = trimethoprim-sulfamethoxazole; A = amoxicillin; AM = amoxicillin clavulanate; S = ampicillin-sulbactam; T = tetracycline; N = nitrofurantoin; G = gentamicin; AK = amikacin; C = chloramphenicol; CX = cefuroxime; CN = cephalexin; CA = ceftazidime; CP = ciprofloxacin;
Red = resistant; Green = intermediate or sensitive.

3.2. Phylogenetic relationships between the isolates and MLST assignation

To study phylogenetic relationships between the *E. coli* isolates, 2,637 orthologous genes shared by all the genomes were identified and aligned using MUSCLE. We also included a selection of reference strains (as “signposts”) deposited in the NCBI. Alignments of the encoded protein sequences were concatenated into a “superstring alignment” comprised of 805,265 amino acid residues, and an NJ phylogenetic tree was constructed (**Figure 1**).

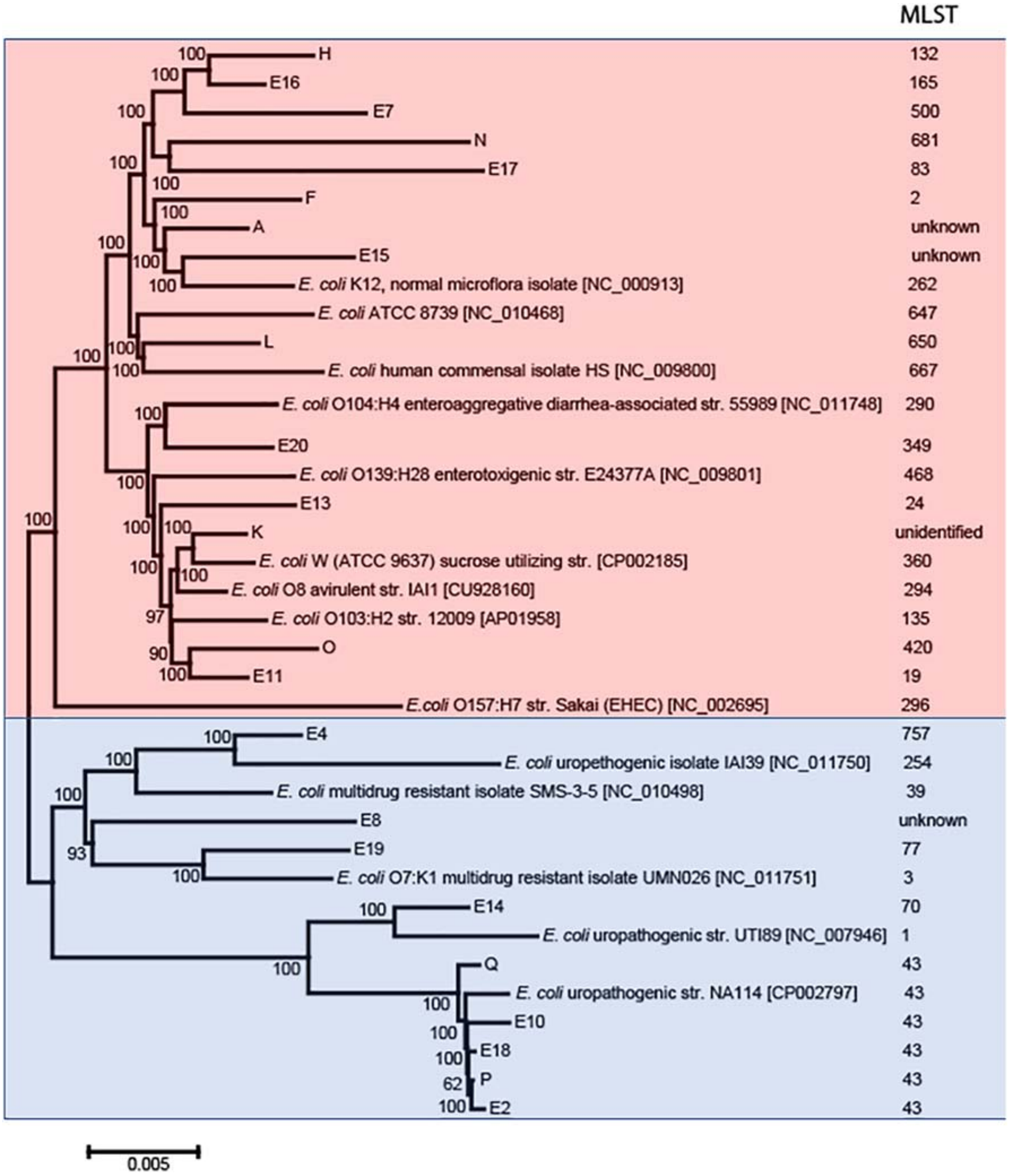

MLST
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
97
90
100
100
100
100
100
93
100
100
100
100
100
100
62
100
H
E16
E7
N
E17
F
A
E15
E. coli K12, normal microflora isolate [NC_000913]
E. coli ATCC 8739 [NC_010468]
L
E. coli human commensal isolate HS [NC_009800]
E. coli O104:H4 enteroaggregative diarrhea-associated str. 55989 [NC_011748]
E20
E. coli O139:H28 enterotoxigenic str. E24377A [NC_009801]
E13
K
E. coli W (ATCC 9637) sucrose utilizing str. [CP002185]
E. coli O8 avirulent str. IAI1 [CU928160]
E. coli O103:H2 str. 12009 [AP01958]
O
E11
E.coli O157:H7 str. Sakai (EHEC) [NC_002695]
E4
E. coli uropethogenic isolate IAI39 [NC_011750]
E. coli multidrug resistant isolate SMS-3-5 [NC_010498]
E8
E19
E. coli O7:K1 multidrug resistant isolate UMN026 [NC_011751]
E14
E. coli uropathogenic str. UTI89 [NC_007946]
Q
E. coli uropathogenic str. NA114 [CP002797]
E10
E18
P
E2
132
165
500
681
83
2
unknown
unknown
262
647
650
667
290
349
468
24
unidentified
360
294
135
420
19
296
757
254
39
unknown
77
3
70
1
43
43
43
43
43
43
0.005

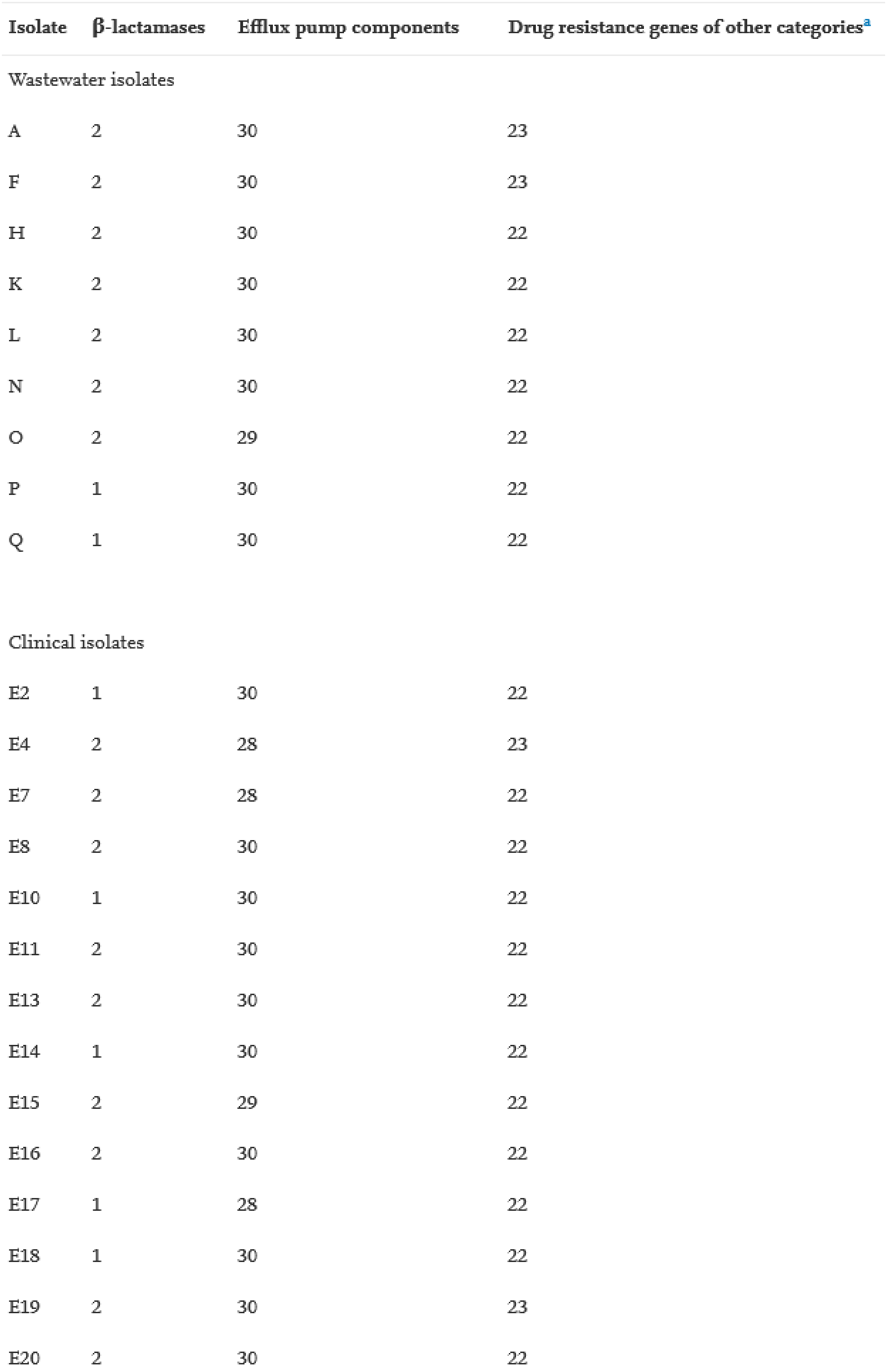
Fig. 1. An NJ phylogenetic tree of clinical and wastewater *E. coli* isolates (sequenced in this study) compared with a selection of NCBI reference strains. Bootstrap numbers are shown at each intermediate split in the tree. The identified MLST of the strains is shown in the corresponding column. The tree divided early on into two broad branches, enriched in either commensal/enterotoxigenic *E. coli* variants (red shading), or uropathogenic isolates (blue shading). [Note that in the wastewater isolate K, the *icdA* was fragmented in the contig preventing definitive assignation of a ST for this strain. However, based on the sequence profile of the other marker genes in isolate K, it may belong either to ST 735, or ST 910.]

The strains segregated in the tree into two broad clades, corresponding to commensal/enterotoxigenic *E. coli* variants (red shading in **Figure 1**), and uropathogenic isolates (blue shading). Although the clinical isolates were roughly equally distributed between these two clusters, most of the wastewater isolates fell into the commensal/enterotoxigenic variant grouping (red shading). The isolates were associated with a diverse range of known MLSTs. Indeed, the only ST represented by more than a single isolate was ST 43, which was associated with a phylogenetic cluster comprising three clinical isolates and two wastewater isolates, as well as the uropathogenic *E. coli* strain NA114. ST 43 is widely distributed around the world, and of the 91 recorded ST 43 strains in the Institute Pasteur *E. coli* MLST database, 9 are clinical uropathogenic isolates. It is also noteworthy that although ST 131 is widely reported as one of the most common clinical *E. coli* isolates worldwide (Pitout and Laupland, 2008; Nicolas-Chanoine et al., 2014), there were no representatives of this ST among the *E. coli* strains we examined. Conversely, our collection also included several strains with STs not yet recorded in the *E. coli* MLST database. For example, the uropathogenic isolate, E8, possessed unique combinations of variants in the sequences of all marker genes except *uidA*, indicating that E8 likely represents a completely new ST of *E. coli*.

3.3. Distribution of antibiotic resistance and virulence-associated genes

The WGS data for all of the isolates indicated that they encode numerous multidrug efflux pumps (including *mdlAB*, *mdtABCD*, *mdtIJ*, *mdtK*, *mdtEF*, *acrAB-acrD-acrEF*, *cmr*, *yddA*, *yojI*, *yjiO*, *emrAB*, *emrD*, *emrKY*, and *hsrA*) and a selection of known antibiotic resistance genes, including *rarD* and *cmlA* for chloramphenicol resistance, and the β-lactamases, *ampH*, *ampC*, *blr*. AMR-genes were searched by the program CARD_RGI looking for similarities of protein sequences with records of it’s own comprehensive database of AMR-proteins (Alcock *et al*., 2020). A summary of the antibiotic resistance genes identified by CARD-RGI in each isolate is shown in **Table 2**. Interestingly, the mean number of all classes of AMR-associated genes was similar in both the wastewater and clinical isolates.

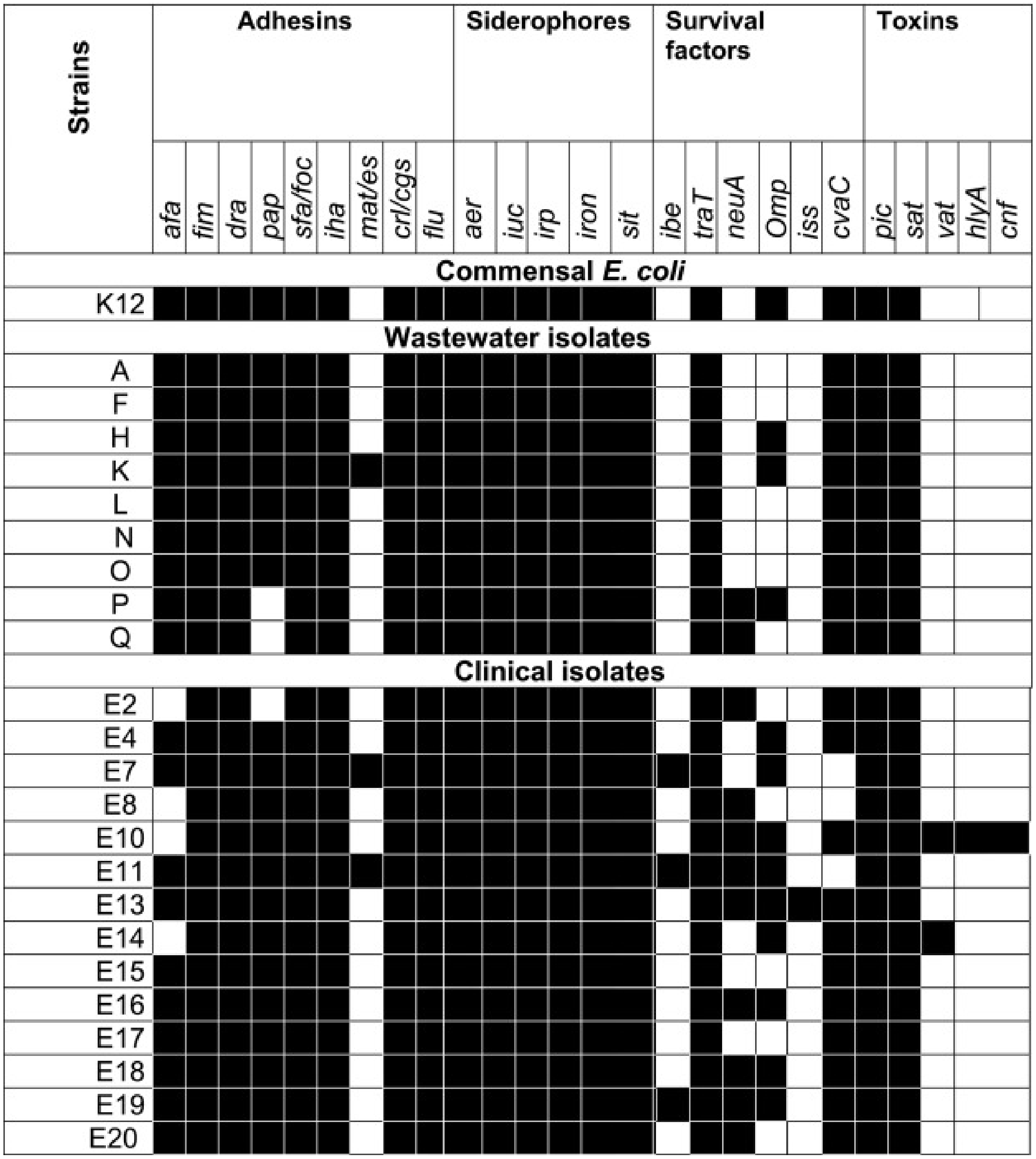
Table 2. Antibiotic resistance gene distribution in the chromosomal DNA of the indicated sequenced isolates. The numbers in each column represent the number of genes encoding the indicated antibiotic resistance mechanism identified in each isolate.



Isolate	β-lactamases	Efflux pump components	Drug resistance genes of other categories[a]
Wastewater isolates			
A	2	30	23
F	2	30	23
H	2	30	22
K	2	30	22
L	2	30	22
N	2	30	22
O	2	29	22
P	1	30	22
Q	1	30	22
Clinical isolates			
E2	1	30	22
E4	2	28	23
E7	2	28	22
E8	2	30	22
E10	1	30	22
E11	2	30	22
E13	2	30	22
E14	1	30	22
E15	2	29	22
E16	2	30	22
E17	1	28	22
E18	1	30	22
E19	2	30	23
E20	2	30	22

[a] This column combines all other types of AMR genes including known transcriptional regulators of drug resistance response, antibiotic-modifying enzymes and uncharacterized drug resistance proteins.

Table 3. Known virulence determinants in the sequenced isolates. Black cells = presence of gene, white cells = absence of gene. BLASTP cutoff values for assignment of gene presence/absence were 0.0001.



Strains	Adhesins									Siderophores					Survival factors						Toxins				
	afa	*fim*	*dra*	*pap*	*sfa/foc*	*iha*	*mat/es*	*crl/cgs*	*flu*	*aer*	*iuc*	*irp*	*iron*	*sit*	*ibe*	*traT*	*neuA*	*Omp*	*iss*	*cvaC*	*pic*	*sat*	*vat*	*hlyA*	*cnf*
Commensal *E. coli*																									
K12	■	■	■	■	■	■		■	■	■	■	■	■	■		■		■		■	■	■			
Wastewater isolates																									
A	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
F	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
H	■	■	■	■	■	■		■	■	■	■	■	■	■		■		■		■	■	■			
K	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■		■	■	■			
L	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
N	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
O	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
P	■	■	■		■	■		■	■	■	■	■	■	■		■	■	■		■	■	■			
Q	■	■	■		■	■		■	■	■	■	■	■	■		■	■			■	■	■			
Clinical isolates																									
E2		■	■		■	■		■	■	■	■	■	■	■		■	■			■	■	■			
E4	■	■	■	■	■	■		■	■	■	■	■	■	■		■		■		■	■	■			
E7	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■			■	■			
E8		■	■	■	■	■		■	■	■	■	■	■	■		■	■				■	■			
E10		■	■	■	■	■		■	■	■	■	■	■	■		■	■	■		■	■	■	■	■	■
E11	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■	■			
E13	■	■	■	■	■	■		■	■	■	■	■	■	■		■	■	■	■	■	■	■			
E14		■	■	■	■	■		■	■	■	■	■	■	■		■		■		■	■	■	■		
E15	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
E16	■	■	■	■	■	■		■	■	■	■	■	■	■		■	■	■		■	■	■			
E17	■	■	■	■	■	■		■	■	■	■	■	■	■		■				■	■	■			
E18	■	■	■	■	■	■		■	■	■	■	■	■	■		■	■	■		■	■	■			
E19	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■		■	■	■			
E20	■	■	■	■	■	■		■	■	■	■	■	■	■		■	■			■	■	■			

The genome sequences of the isolates also contained multiple genes known to be associated with virulence. Sarowska *et al*. (2019) listed several key groups of virulence factors frequently found in pathogenic *E. coli* isolates, including several classes of adhesins, siderophores, toxins and proteins important for intracellular survival of pathogens, colonization of non-GI tract tissues, and immune system avoidance. A summary of distribution of virulence-associated determinants in the isolates studied here is shown in **Table 3**. The data indicate an enrichment of genes associated with survival in the host among the clinical isolates including virulence-associated siderophores of five classes, *aer*, *iuc*, *irp*, *iron* and *sit*, which were present on the chromosomes of all the sequenced isolates. In several cases, additional copies of these siderophore genes were also found on plasmids (see **Figure 3** and discussion below).

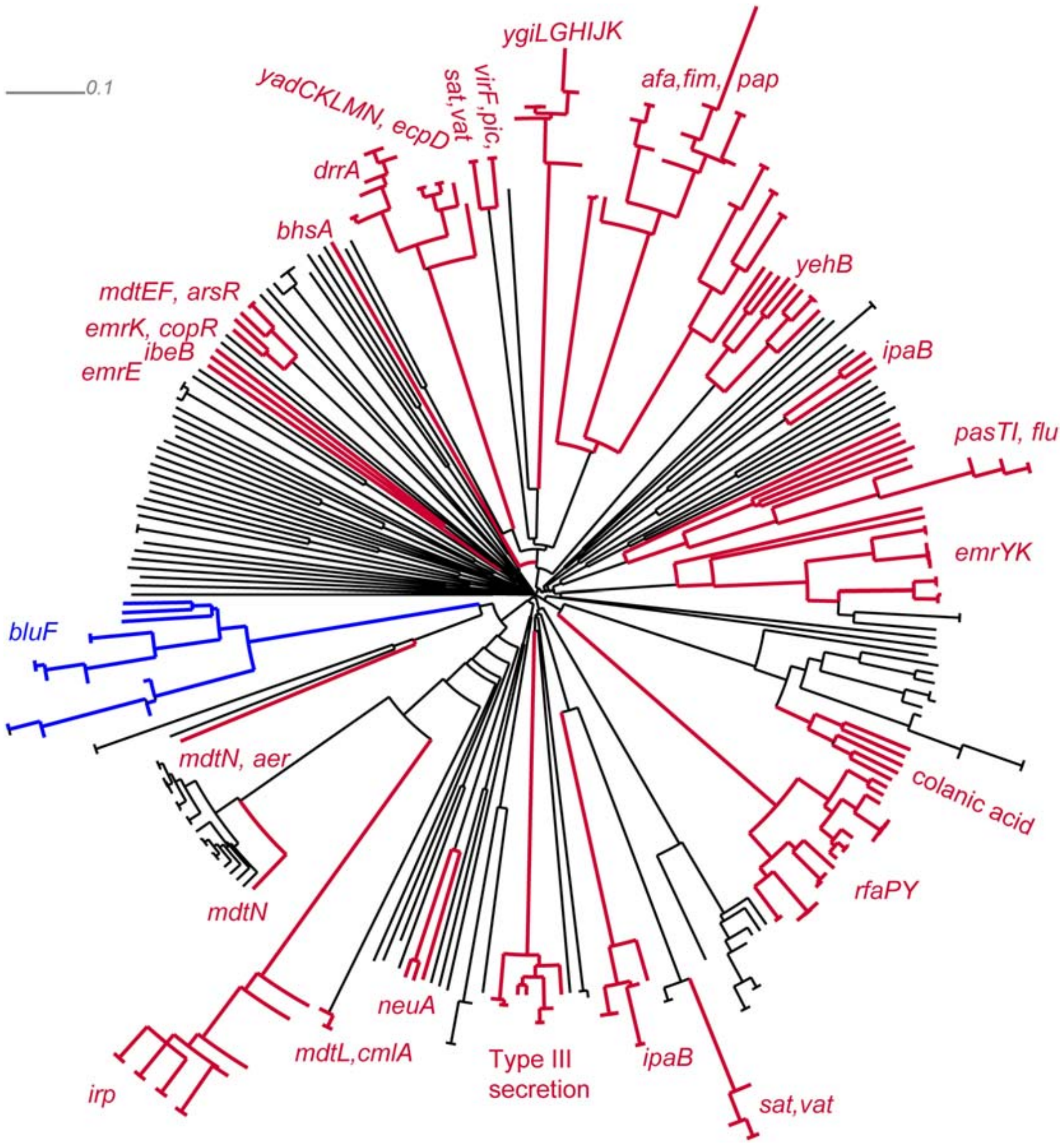


Fig. 2. Clustering of GIs based on gene content using an NJ algorithm (in turn, based on the distance matrix given in Eq. 1). A single line radiating from the center to the edge of the cladogram indicates a GI in a specific isolate. Branched radiating lines indicate diversification of a GI present in multiple isolates. The clustered GIs were further interrogated to identify AMR and virulence-associated genes using BLASTP. Groups of GIs containing known AMR and virulence-associated genes are shown red. GIs with the blue light- and temperature-regulated genes are shown blue.

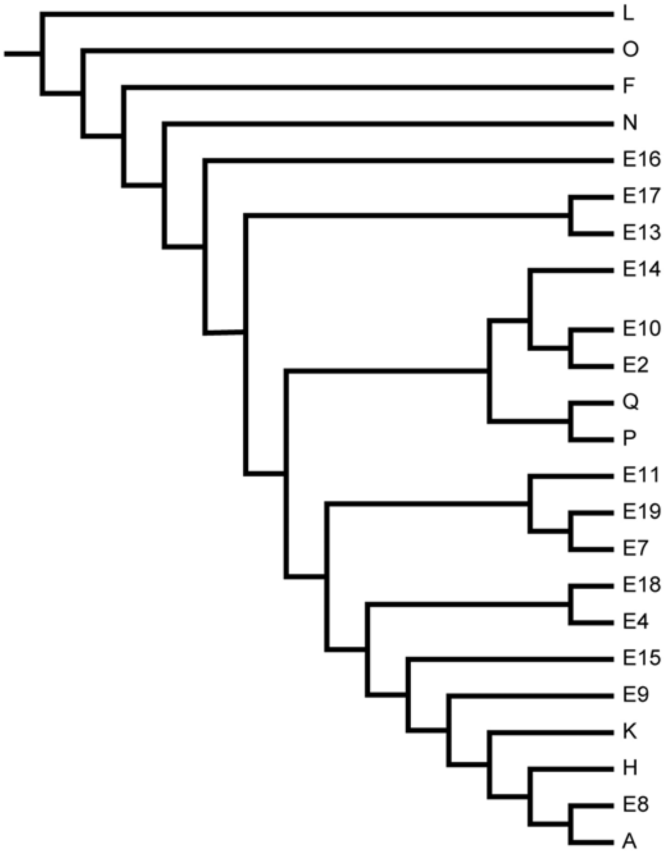
3.4. Genomic islands and the distribution of drug resistance and virulence-associated genes

We next examined the genomic context of the AMR and virulence-associated (VA) genes i.e., are these genes located in the conserved “core chromosome”, or are they associated with the variable “accessory” genome (GIs and plasmids)? The genome sequences of all the isolates contained multiple horizontally-acquired GIs, including prophages, transposable elements and integrons. The GIs were clustered based on the “shared gene” algorithm in Eqn 1. We then further interrogated each resulting GI cluster for its AMR and VA gene content (**Figure 2**).

Although the gene content of GIs was generally highly variable, this clustering approach revealed some intriguing genetic conservation. In the left-top corner of the cladogram (**Figure 2**), there is a clutch of individual GIs bearing several VA genes. For example, the small multidrug resistance efflux transporter *emrE* is located on a GI found only in environmental isolate F, whereas the cell invasion gene *ibeB*, along with several genes conferring resistance to copper (collectively named “*copR”* in the figure) and a multidrug resistance gene (*emrK*) were found in two GIs from isolate E19. Clockwise from these, there is a cluster of five highly similar GIs found in isolates F, K, L, E17 and E20, which bear genes for arsenate resistance (collectively named *arsR* in the figure) and the *mdtEF* multidrug efflux operon. Another individual GI found in isolate E9 contains the multiple stress resistance gene, *bhsA*. A large group of 13 GIs from both clinical and environmental isolates contains an operon encoding fimbria-like genes (*yadCKLMN* and *ecpD*), which are known to function as virulence-associated adhesins in enterohaemorrhagic and uropathogenic *E. coli* (Spurbeck et al., 2011; Chingcuanco et al., 2012; Stacy et al., 2014). Within this group, seven GIs from isolates N, E7, E9, E13, E14, E16 and E19 form a distinct sub-cluster possessing the daunorubicin/ doxorubicin resistance-associated gene, *drrA*. Four GIs from the genome sequences of isolates P, Q, E2 and E10 contain virulence-associated toxin-encoding genes, including *sat*, *vat* and *pic*, together with a virulence regulon transcriptional activator-encoding gene, *virF*. Eight GIs from isolates K, L, E4, E7, E11, E16 and E20 contain the *ygiLGHIJK* operon, which encodes another set of virulence-associated fimbria-like proteins (Spurbeck et al., 2011). A very large group of 23 GIs, which were shared many of the sequenced isolates, comprises additional fimbria-encoding genes, *afa*, *fim* and *pap*. The *E. coli* fimbrial protein YehB is responsible for adhesion to abiotic surfaces (Ravan and Amandadi, 2015) and as such may contribute to the dissemination of nosocomial infections in hospitals, and possibly, also survival in the environment. This gene was found in eight GIs from isolates F, K, O, E4, E11, E13, E19 and E20. Another adhesion gene, *flu* (*agn43*), in combination with a persistence and stress-resistance toxin-antitoxin system, *pasTI*, was present in seven GIs from isolates H, E7, E11, E14, E16, E19 and E20. Twelve GIs from isolates H, O, K, E7, E8, E11, E13, E14, E15, E16, E17 and E20 carry the multidrug resistance operon, *emrYK*, and six GIs from environmental isolates A, H, N, O, P and Q encode a colanic acid biosynthetic pathway. Colanic acid is a capsular carbohydrate of uropathogenic bacteria and is known to be important for biofilm formation (Prigent-Combaret et al., 2000; Hanna et al., 2003). A short operon, *rfaPY*, which encodes a pair of lipopolysaccharide core heptose kinases (I and II), was found in 21 GIs. These genes may encode potential virulence factors, since the activation of heptose precursors is used in the biosynthesis of lipopolysaccharides (LPS), and LPS is known to contribute towards the pathogenicity of enterotoxigenic *E. coli* (Maigaard Hermansen et al., 2018). Interestingly, orthologues of the *sat* and *vat* toxins mentioned above are encoded on another group of four GIs from isolates E10, E11, E14 and E18. We note that isolate E10 also encodes *sat* and *vat* on a different GI. An increased copy number of these genes may influence the virulence of this strain (Elliot et al., 2013; Slager and Veening, 2016). Similarly, we also noted that the invasin, *ipaB*, was distributed between nine GIs which segregated into two distinct clusters (Venkatesan et al., 1988). The smaller of these clusters contains GIs from isolates E4, E8 and E19, whereas the larger one comprises GIs from E2, E8, E10, E14, E18 and E19. Once again, we noted that two of the isolates (E8 and E19) carried paralogous copies of *ipaB* in two separate GIs. Nine GIs from isolates H, K, L, E7, E8, E13, E16, E19 and E20 carry a large operon encoding a type III secretion system (Tree et al., 2009). The anti-phagocytosis factor-encoding gene, *neuA*, was found in three related

GIs from isolates P, Q and E17. Three GIs from isolates E4, E8 and E19 carried a multidrug resistance gene, *mdtL*, and a chloramphenicol resistance gene *cmlA*. A group of 12 large GIs contained a polypeptide synthase gene (*irp*) encoding the siderophore, yersiniabactin. Host organisms often sequester iron to inhibit the growth of pathogens. Consequently, iron acquisition systems are often considered to be virulence factor/ survival mechanisms associated with pathogenicity (Skaar 2010). Several other virulence-associated siderophores are involved in iron scavenging and transportation. The genes encoding the synthesis and transport of these siderophores were abundant in the isolates (**Table 2**). Orthologues of the multidrug resistance-associated gene, *mdtN*, were also found in two loosely clustered GIs from isolates E8 and E9. The same GI from E9 carries also another siderophore (aerobactin) encoding gene, *aer*. Finally, a group of 16 GIs contains several regulatory genes including the blue light- and temperature-regulated anti-repressor *bluF*. BluF is light and temperature sensing protein, which regulates biofilm formation (Tschowri et al., 2009). While it is not considered as a virulence factor, it may influence the survival of bacteria in the environment. These GIs were found in both the environmental and clinical samples (specifically, isolates A, F, L, P, Q, E2, E4, E7, E10, E11, E14, E16, E17, E19 and E20). Notably, in the genome of isolate E2, this GI was duplicated.

Not all virulence genes were associated with specific clusters of GIs. For example, the major fimbrial subunit gene *lpfA*, which influences epithelial cell invasion by mastitis-associated *E. coli* (Dogan et al., 2012), was found in seven unrelated GIs from clinical isolates E4, E9, E11, E13, E14, E15 and E20. We also noted that many GIs were enriched in genes encoding phage-related proteins and integrases (suggesting the likely mode of horizontal transmission), and many others also encoded metabolic enzymes and transmembrane transporters. It is possible that these cargo genes are involved in adaptation to specific environments, and may indirectly impact on virulence, AMR or survival in the face of environmental challenges.



	AMR	VA
L	**1** BL:0, EP:0, Other:1	No
O	No	*sit,pic,sat,vat*
F	No	No
N	No	*sit,pic,sat,vat*
E16	**9** BL:1, EP:2, Other:6	*afa,dra,sfa,pic,sat,vat*
E17	**13** BL:1, EP:1, Other:11	*iha,sit*
E13	**7** BL:1, EP:1, Other:5	*iha,sit*
E14	**1** BL:0, EP:0, Other:1	*afa,dra,fim,hly,pap, sfa,foc,iha,iron,sit*
E10	**9** BL:1, EP:0, Other:8	*iha,iron,sit*
E2	**6** BL:0, EP:0, Other:6	*iha,iron,sit*
Q	**4** BL:0, EP:0, Other:4	*sit*
P	**4** BL:0, EP:0, Other:4	*sit*
E11	**10** BL:2, EP:1, Other:9	*aer,mat,traT*
E19	**4** BL:1, EP:1, Other:2	*iha,sit,traT,pic,vat*
E7	**3** BL:0, EP:0, Other:3	*iha,iron,sit,traT*
E18	**12** BL:1, EP:1, Other:10	*iha,sit*
E4	**1** BL:0, EP:0, Other:1	*afa,dra,sfa*
E15	**5** BL:1, EP:2, Other:2	*iha,sit,pic,sat,vat*
E9	**4** BL:0, EP:3, Other:1	*irp,sit*
K	**3** BL:0, EP:0, Other:3	No
H	**1** BL:0, EP:0, Other:1	No
E8	**1** BL:0, EP:0, Other:1	No
A	**1** BL:0, EP:0, Other:1	No

Fig. 3. Dollo parsimony clustering of plasmid contigs by presence/absence of shared homologous genes identified by a reciprocal BLASTP alignment. In line with the end-node titles, total numbers of antimicrobial resistance (AMR) genes are shown in bold followed by numbers per categories of β-lactamases (BL), efflux pumps (EP) and other categories of drug resistance genes. A list of the encoded virulence-associated (VA) genes is shown in the adjacent column. VA genes found on the plasmid-born contigs are: *aer* – siderophore; *afa* – Afa-like afimbrial adhesins; *dra* – Dra-like surface-exposed fimbria associated proteins common in uropathogenic *E. coli*; *fim* – type I fimbria proteins; *foc* – Foc-like adhesins to intestinal epithelial cells; *hly* – hemolysin transport protein creating pores in host cell membranes; *iha* – iron-regulated adhesin; *iron* – siderophore receptor; *irp* – yersiniabactin siderophore synthesis protein; *mat* – meningitis associated and temperature regulated fimbriae; *pap* – pilin, colonization factor in extraintestinal infections stimulating the production of cytokines by T lymphocytes; *pic* – mucin degrading serine protease; *sat* – serine protease autotransporter toxin; *sfa* – Sfa-like adhesins to intestinal epithelial cells; *sit* – Sit-like iron transmembrane transporters; *traT* – phagocytosis inhibitor; *vat* – vacuolating autotransporter toxin.

3.5. Role of plasmids in distribution of drug resistance and virulence associated genes

Predicted plasmid-borne contigs were identified in the assemblies from all of the isolates. These contigs contained between 4 coding sequences (CDS) in E8 up to 288 CDS in E11. It should be noted that assembly of plasmid contigs is problematic due to a higher level of sequence variability in these regions. Therefore, we note that the obtained contigs may not represent whole sequences of the plasmids in the studied isolates. Grouping of concatenated plasmid-born contigs from different isolates based on shared homologous genes is shown in **Figure 3**. This figure also summarizes the numbers of AMR and VA genes found in these contigs.

It is immediately clear from inspection of the data in **Figure 3** that the plasmid-borne DNA from clinical isolates is enriched in AMR and VA genes compared with the wastewater isolates. Indeed, the plasmids from environmental isolates A, H, K, F, and L encoded no virulence-associated genes and only one or no AMR determinants. Exceptions to this trend included plasmids from the environmental isolates P and Q, which contained multiple AMR and VA genes, whereas the largest plasmid-associated contig in the dataset (from clinical isolate E8) contained no AMR or virulence determinants. The largest number of AMR determinants was identified in the plasmid DNA from E17, E18, E11 and E9, with the highest number of virulence-associated genes found in the plasmid from isolate E14. Interestingly, the plasmid-borne virulence-associated genes were often duplicates of genes already present on the chromosome, whereas the plasmid-borne AMR determinants were more specific.

4. DISCUSSION

In this study, we found that clinical and wastewater isolates possessed similar overall numbers of chromosomally encoded AMR and VA genes, irrespective of their origin (**Table 2** and **Table 3**). However, the clinical isolates displayed an enrichment of AMR and VA genes in their plasmidome. Of note, a number of these genes were present in multiple paralogous copies (either on the chromosome, in GIs, or on plasmids). Gene duplication is a key driver of functional diversification and is also a facile means of increasing gene expression through increased copy number (Elliot et al., 2013; Slager and Veening, 2016). In bacteria, there are several examples of genes encoding basic metabolic functions being present in two copies, with one copy on the chromosome and the other on a plasmid (Zheng et al., 2015). Our observation, that plasmids from the clinical isolates were enriched in AMR and VA genes indicates that these mobile elements may play a key role in pathogenicity. The plasmids associated with environmental isolates P and Q were exceptions to this general trend, since these were enriched with AMR and virulence-associated genes. Isolates P and Q are notable since they are phylogenetically related to the clinical isolates E2, E10 and E18, and all five isolates belonging to the uropathogenic ST 43 (**Figure 1**). There may be two non-exclusive explanations for this observation. First, P and Q may be disseminated "clinical isolates" that just happen to have been captured in the local watershed following e.g., human discharge activities. Alternatively, these particular plasmids may confer a fitness advantage in the wastewater environment. The factors increasing the survival of these uropathogenic strains may include the presence of colanic acid biosynthetic genes and the light- and temperature-sensing anti-repressor, *bluF*. These genes have been shown to play an important role in the modulation of biofilm formation (Prigent-Combaret et al., 2000; Hanna et al., 2003; Tschowri et al., 2009) and therefore confer a potential advantage to ST 43 for survival in both non-host and host environments.

Collectively, our data suggest that plasmid-borne functions confer an advantage to *E. coli* in terms of infection and/or withstanding exposure to the antimicrobials commonly used to treat these infections. At the same time, these genes presumably confer a significant burden on bacterial growth and replicative potential during transit through the environment. Consequently, when not infecting a host, there is likely a selection pressure on bacterial populations to lose these plasmid-borne genes. This presumably generates a drive towards redistributing the encoded functions from the plasmid to the chromosome, which may be another reason why we see apparently multiple paralogous copies of some genes (Andersson and Hughes, 2010; Kussell, 2013; Melnyk et al., 2015).

In conclusion, our data confirm and extend previous reports indicating the high genomic diversity of *E. coli* in humans from tropical areas (Escobar-Páramoet al 2004; Richter et al., 2018), and furthermore, implicate plasmids as a key driver of AMR and VA gene dissemination, especially among clinical isolates. Although plasmids come with an intrinsic fitness cost associated with their replication, we noted that in some circumstances (e.g., the isolates of ST 43) plasmid-borne AMR and VA genes were also associated with the environmental isolates. This ability to maintain plasmid-borne functions outside the host may contribute towards the global success of this sequence type.

Research data access

All sequence data generated in this study have been submitted to the NCBI BioProject database (BioProject: https://www.ncbi.nlm.nih.gov/) under accession numbers: PRJNA606749; PRJNA630767; PRJNA630874; PRJNA606697.

Funding and acknowledgements

This work was supported by the Cambridge – Africa ALBORADA Research Fund and the Academy of Medical Sciences GCRF networking grant reference GCRFNG/100095. TJO was supported by a studentship (NC/P001564/1) from the NC3Rs awarded to MW. WF was supported by a studentship from the Cambridge Trust-CONACyT. This study was approved by Jaramogi Oginga Odinga Referral & Teaching Hospital IREC; reference number ERC.IB/VOL.1/141.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest.

Author contributions

SAW, WAS, conceptualization, investigation and original draft; TJO, WF, VD investigation and methodology; OR bioinformatics analysis and original draft; MW supervision, draft review, funding procurement and editing.

References

Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh. W., Nguyen, A.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., Hernandez-Koutoucheva, A., Sharma, A.N., Bordeleau, E., Pawlowski, A.C., Zubyk, H.L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G.L., Beiko, R.G., Brinkman, F.S.L., Hsiao, W.W.L., Domselaar, G.V., McArthur, A.G. 2020. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 48(D1):D517-D525. doi: 10.1093/nar/gkz935

Andersson, D.I., Hughes, D. 2010. Antibiotic resistance and its cost: is it possible to reverse resistance? Nat Rev Microbiol. 8:260–271. doi: 10.1038/nrmicro2319

Arredondo-Alonso, S. Rogers, M.R.C., Braat, J.C., Verschuuren, T.D., Top, J., Corander, J., Willems, R.J.L., Schürch, A.C. 2018. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microb Genom. 4 (11). https://doi.org/10.1099/mgen.0.000224

Besser, J., Carleton, H.A., Gerner-Smidt, P., Lindsey, R.L., Trees, E. 2018. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical Microbiol. Infect. 24:335-341. doi: 10.1016/j.cmi.2017.10.013

Bezuidt, O., Lima-Mendez, G., Reva, O.N. 2009. SEQWord Gene Island Sniffer: a program to study the lateral genetic exchange among bacteria. World Acad. Sci. Eng. Technol. 58, 1169-1174.

Blyton, M.D., Cornall, S.J., Kennedy, K., Colligon, P., Gordon, D.M. 2014. Sex-dependent competitive dominance of phylogenetic group B2 *Escherichia coli* strains within human hosts. Environ Microbiol Rep. 6(6):605–10. doi: 10.1111/1758-2229.12168

Cao, Y., Fanning, S., Proos, S., Jordan, K., Srikumar, S. 2017. A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Front Microbiol. 8:1829. https://doi.org/10.3389/fmicb.2017.01829

Chingcuanco, F., Yu, Y., Kus, J.V., Que, L., Lackraj, T., Lévesque, C.M., Foster, B.D. 2012. Identification of a novel adhesin involved in acid-induced adhesion of enterohaemorrhagic *Escherichia coli* O157:H7. Microbiol. 158:2399-2407. doi:10.1099/mic.0.056374-0.

Clermont, O., Bonacorsi, S., Bingen, E. 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66:4555–4558. doi: 10.1128/aem.66.10.4555-4558.2000.

Conceição, R.A., Ludovico, M.S., Andrade, C.G.T.J., Yano, T. 2012. Human sepsis-associated *Escherichia coli* (SEPEC) is able to adhere to and invade kidney epithelial cells in culture. Braz J Med Biol Res. 2012(45):417–424. doi: 10.1590/S0100-879X2012007500057.

CLSI. 2017. Performance standard for antimicrobial susceptibility testing 27th ed. CLSI supplement M100 Wayne, PA: Clinical and Laboratory Standards Institutes.

Dautin, N. 2010. Serine protease autotransporters of *Enterobacteriaceae* (SPATEs): biogenesis and function. Toxins (Basel). 2(6):1179-1206. doi:10.3390/toxins2061179.

Dogan, B., Rishniw, M., Bruant, G., Harel, J., Schukken, Y.H., Simpson, K.W. 2012. Phylogroup and *lpfA* influence epithelial invasion by mastitis associated *Escherichia coli*. Vet Microbiol. 159(1-2):163-170. doi: 10.1016/j.vetmic.2012.03.033

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792-1797. https://doi.org/10.1093/nar/gkh340

Elliott, K.T., Cuff, L.E., Neidle, E.L. 2013. Copy number change: evolving views on gene amplification. Future Microbiol. 8(7):887-899. doi: 10.2217/fmb.13.53

Emms, D.M., Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157. https://doi.org/10.1186/s13059-015-0721-2

European Centre for Disease Prevention and Control. 2018. Monitoring the use of whole-genome sequencing in infectious disease surveillance in Europe. Stockholm: ECDC.

Escobar-Páramo, P., Grenet, K., Le Menac'h, A., Rode, L., Salgado, E., Amorin, C., Gouriou, S., Picard, B., Rahimy, M.C., Andremont, A., Denamur, E., Ruimy, R. 2004. Large-scale population structure of human commensal *Escherichia coli* isolates. Appl Environ Microbiol. 70(9):5698-5700. doi: 10.1128/AEM.70.9.5698-5700.2004

Hanna, A., Berg, M., Stout, V., Razatos, A. 2003. Role of capsular colanic acid in adhesion of uropathogenic *Escherichia coli*. Appl Environ Microbiol. 69(8):4474-4481. doi: 10.1128/AEM.69.8.4474-4481.2003

Huson, D.H., Steel, M. 2004. Phylogenetic trees based on gene content. Bioinformatics. 20(13):2044-2049. doi: 10.1093/bioinformatics/bth198

Huson, D.H., Scornavacca, C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 61(6):1061-1067. doi: 10.1093/sysbio/sys062

Jafari, A., Aslani, M.M., Bouzari, S. 2012*. Escherichia coli*: A brief review of diarrheagenic pathotypes and their role in diarrheal diseases in Iran. Iran. J. Microbiol. 4 (3), 102–117. PMCID: PMC3465535

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol. 35(6):1547-1549. doi: 10.1093/molbev/msy096

Kussell, E. 2013. Evolution in microbes. Annu Rev Biophys. 42:493–514. doi: 10.1146/annurev-biophys-083012-130320

Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., Lund, O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 50(4):1355-1361. doi: 10.1128/JCM.06094-11

Maigaard Hermansen, G.M., Boysen, A., Krogh, T.J., Nawrocki, A., Jelsbak, L., Møller-Jensen, J. 2018. HldE is important for virulence phenotypes in enterotoxigenic *Escherichia coli*. Front Cell Infect Microbiol. 8: 253. doi: 10.3389/fcimb.2018.00253

Malaho, C., Sifuna, A.W., Shivoga, A.W. 2018. Antimicrobial resistance patterns of *Enterobacteriaceae* recovered from wastewater, sludge and dumpsite environments in

Kakamega town, Kenya; Afr. J. Microbiol. Res. 12(28), pp. 673-680. https://doi.org/10.5897/AJMR2017.8656

Melnyk, A.H., Wong, A., Kassen, R. 2015. The fitness costs of antibiotic resistance mutations. Evol Appl. 8:273–283. doi: 10.1111/eva.12196

Nicolas-Chanoine, M.H., Bertrand, X., Madec, J.Y. 2014. *Escherichia coli* ST131, an intriguing clonal group. Clin Microbiol Rev. *27*(3), 543–574. https://doi.org/10.1128/CMR.00125-13

Pitout JD, Laupland KB. 2008. Extended-spectrum beta-lactamase-producing *Enterobacteriaceae:* an emerging public-health concern. Lancet Infect Dis. 8(3):159–66.doi: 10.1016/S1473-3099(08)70041-0

Prigent-Combaret, C., Prensier, G., Le Thi, T.T., Vidal, O., Lejeune, P., Dorel, C. 2000. Developmental pathway for biofilm formation in curli-producing *Escherichia coli* strains: role of flagella, curli and colanic acid. Environ Microbiol. 2(4):450-464. doi: 10.1046/j.1462-2920.2000.00128.x

Rantsiou, K., Kathariou, S., Winkler, A., Skandamis, P., Saint-Cyr, M.J., Rouzeau-Szynalski, K., Amézquita, A. 2018. Next generation microbiological risk assessment: opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. Int. J. Food Microbiol. 287:3-9. doi: 10.1016/j.ijfoodmicro.2017.11.007

Ravan, H., Amandadi, M. 2015. Analysis of *yeh* fimbrial gene cluster in *Escherichia coli* O157:H7 in order to find a genetic marker for this serotype. Curr Microbiol. 71(2):274-282. doi: 10.1016/j.ijfoodmicro.2017.11.007

Richter, T.K.S., Hazen, T.H., Lam, D., Coles, C.L., Seidman, J.C., You, Y., Silbergeld, E.K., Fraser, C.M., Rasko, D.A. 2018. Temporal variability of *Escherichia coli* diversity in the gastrointestinal tracts of Tanzanian children with and without exposure to antibiotics. mSphere. 3 (6): 3:e00558-18. doi: 10.1128/mSphere.00558-18

Sarowska, J., Futoma-Koloch, B., Jama-Kmiecik, A., Frej-Madrzak, M., Ksiazczyk, M., Bugla-Ploskonska, G., Choroszy-Krol, I. 2019. Virulence factors, prevalence and potential transmission of extraintestinal pathogenic *Escherichia coli* isolated from different sources: recent reports. Gut Pathog. 11:10. doi: 10.1186/s13099-019-0290-0

Slager, J., Veening, J.W. 2016. Hard-Wired Control of Bacterial Processes by Chromosomal Gene Location. Trends Microbiol. 24(10):788-800. doi: 10.1016/j.tim.2016.06.003

Spurbeck, R.R., Stapleton, A.E., Johnson, J.R., Walk, S.T., Hooton, T.M., Mobley, H.L. (2011). Fimbrial profiles predict virulence of uropathogenic *Escherichia coli* strains: contribution of *ygi* and *yad* fimbriae. Infect Immun. 79(12):4753-4763. doi: 10.1128/IAI.05621-11

Skaar, E.P. 2010. The battle for iron between bacterial pathogens and their vertebrate hosts. PLoS Pathog. 6(8):e1000949. https://doi.org/10.1371/journal.ppat.1000949

Stacy, A.K., Mitchell, N.M., Maddux, J.T., De la Cruz, M.A., Durán, L., Girón, J.A., Curtiss, R 3rd., Mellata, M. 2014. Evaluation of the prevalence and production of *Escherichia coli* common pilus among avian pathogenic *E. coli* and its role in virulence. PLoS One. 9(1):e86565. doi: 10.1371/journal.pone.0086565

Talavera, G., Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. 56:564-577. doi: 10.1080/10635150701472164

Tree, J.J., Wolfson, E.B., Wang, D., Roe, A.J., Gally, D.L. 2009. Controlling injection: regulation of type III secretion in enterohaemorrhagic *Escherichia coli*. Trends Microbiol. 17(8):361-370. doi: 10.1016/j.tim.2009.06.001

Tschowri, N., Busse, S., Hengge, R. 2009. The BLUF-EAL protein YcgF acts as a direct anti-repressor in a blue-light response of *Escherichia coli*. Genes Dev. 23(4):522-534. doi: 10.1101/gad.499409

Venkatesan, M., Buysse, J.M., Vandendries, E., Kopecko, D.J. (1988). Development and testing of invasion-associated DNA probes for detection of *Shigella* spp. and enteroinvasive *Escherichia coli*. J Clin Microbiol. 26(2):261-266. PMCID: PMC266263

Vernet, G., Mary, C., Altmann, D.M., Doumbo, O., Morpeth, S., Bhutta, Z.A., Klugman, K.P. 2014. Surveillance for Antimicrobial Drug Resistance in Under-Resourced Countries; Emerg Infect Dis. 20(3):434-441. https://dx.doi.org/10.3201/eid2003.121157

Wood, D.E., Salzberg, S.L. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46

Zheng, J., Guan, Z., Cao, S., Peng, D., Ruan, L., Jiang, D., Sun, M. 2015. Plasmids are vectors for redundant chromosomal genes in the *Bacillus cereus* group. BMC Genomics 16(1):6. doi: 10.1186/s12864-014-1206-5

Zhou, Z., Alikhan, N.F., Mohamed, K., Fan, Y., Brown, D., Chattaway, M., Dallman, T., Delahay, R., Kornschober, C., Pietzka, A., Malorny, B., Petrovska, L., Davies, R., Robertson, A., Tyne, W., Weill, F.X., Accou-Demartin, M., Williams, N., Achtman, M. 2020. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. Genome Res. 30(1):138-152. doi: 10.1101/gr.251678.119