frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 06 January 2017
doi: 10.3389/fcimb.2016.00200

Chaperone-Usher Pili Loci of Colonization Factor-Negative Human Enterotoxigenic *Escherichia coli*

Felipe Del Canto[1], Miguel O'Ryan[1], Mirka Pardo[1], Alexia Torres[1], Daniela Gutiérrez[1], Leandro Cádiz[1], Raul Valdés[1,2], Aquiles Mansilla[1], Rodrigo Martínez[1], Daniela Hernández[1], Benjamin Caro[1], Myron M. Levine[3], David A. Rasko[4], Christopher M. Hill[5], Mihai Pop[5], O. Colin Stine[6] and Roberto Vidal[1]**

[1] Programa de Microbiología y Micología, Instituto de Ciencias Biomédicas, Facultad de Medicina, Universidad de Chile, Santiago, Chile, [2] Facultad de Química y Biología, Universidad de Santiago de Chile, Santiago, Chile, [3] Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, MD, USA, [4] Department of Microbiology and Immunology, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA, [5] Center for Bioinformatics and Computational Biology, University of Maryland Institute for Advanced Computer Studies, College Park, MD, USA, [6] Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA

OPEN ACCESS

Edited by:
Ariel Blocker,
University of Bristol, UK

Reviewed by:
Guoqiang Zhu,
Yangzhou University, China
Anne-Marie Krachler,
University of Texas Health Science Center at Houston, USA

****Correspondence:***
Felipe Del Canto
felipedelcanto@med.uchile.cl
Roberto Vidal
rvidal@med.uchile.cl

Received: *25 August 2016*
Accepted: *14 December 2016*
Published: *06 January 2017*

Citation:
Del Canto F, O'Ryan M, Pardo M, Torres A, Gutiérrez D, Cádiz L, Valdés R, Mansilla A, Martínez R, Hernández D, Caro B, Levine MM, Rasko DA, Hill CM, Pop M, Stine OC and Vidal R (2017) Chaperone-Usher Pili Loci of Colonization Factor-Negative Human Enterotoxigenic Escherichia coli. Front. Cell. Infect. Microbiol. 6:200. doi: 10.3389/fcimb.2016.00200

Enterotoxigenic *Escherichia coli* (ETEC) is one of the most common causes of diarrhea worldwide. Among the 25 different ETEC adhesins, 22 are known as "colonization factors" (CFs), of which 17 are assembled by the chaperone-usher (CU) mechanism. Currently, there is no preventive therapy against ETEC, and CFs have been proposed as components for vaccine development. However, studies of diarrhea-causing ETEC strains worldwide indicate that between 15 and 50% of these are negative for known CFs, hindering the selection of the most widespread structures and suggesting that unknown adhesins remain to be identified. Here, we report the result of a comprehensive analysis of 35 draft genomes of ETEC strains which do not carry known adhesin genes; our goal was to find new CU pili loci. The phylogenetic profiles and serogroups of these strains were highly diverse, a majority of which produced only the heat-labile toxin. We identified 10 pili loci belonging to CU families β (1 locus), γ_2 (7 loci), κ (1 locus), and π (1 locus), all of which contained the required number of open reading frames (ORFs) to encode functional structures. Three loci were variants of previously-known clusters, three had been only-partially described, and four are novel loci. Intra-loci genetic variability identified would allow the synthesis of up to 14 different structures. Clusters of putative γ_2-CU pili were most common (23 strains), followed by putative β-CU pili (12 strains), which have not yet been fully characterized. Overall, our findings significantly increase the number of ETEC adhesion genes associated with human infections.

Keywords: ETEC, adhesin, adhesin negative-ETEC, colonization factors, chaperone-usher pili, genome analysis

INTRODUCTION

Enterotoxigenic *Escherichia coli* (ETEC), which includes a wide diversity of strains, are one of six categories of diarrheagenic *E. coli* (Croxen et al., 2013). These ETEC are a common cause of watery diarrhea worldwide, primarily affecting children living in resource-poor settings of developing countries and travelers who visit these endemic regions (Qadri et al., 2005). Humans are the natural reservoir for ETEC; transmission is associated with consumption of food or water contaminated with human feces (Qadri et al., 2005). After infecting the small bowel epithelium, ETEC induce electrolyte and water loss by producing at least one of two enterotoxins, which distinguish them from other diarrheagenic *E. coli*: a heat stable toxin (ST) and/or a heat labile toxin (LT) (Turner et al., 2006). Strains that infect humans can produce two different types of STs, STh (human variant), and STp (pig variant), but only one type of LT (LT-I) (Croxen et al., 2013). An essential part of infection is attachment to intestinal cells, which ETEC accomplish using a diverse array of adhesins; among these, colonization factors (CFs) have been the most studied to date (Croxen et al., 2013).

Currently, no effective preventive therapy against ETEC is available. Vaccines could be a feasible alternative to reduce the associated morbidity and mortality, particularly in resource-poor settings in developing countries (Ahmed et al., 2013). Epidemiological studies, including characterization of ETEC isolates worldwide, have been the basis for selection of the most widespread antigens for vaccine development. Both enterotoxins and ETEC adhesins have been proposed and tested as components for vaccine candidates (Isidean et al., 2011; O'Ryan et al., 2015; Zhang and Sack, 2015). Thus, CFs CFA/I, CS1, CS2, CS3, CS4, CS5, CS6, and CS7, or their components, have been included in several formulations, however none of these candidates have passed phase III trials (O'Ryan et al., 2015; Zhang and Sack, 2015).

Although 22 different CFs and three other non-pili adhesins (Tia, TibA, and EtpA) have been identified in ETEC strains, a key obstacle to building an effective vaccine based on these antigens is the significant proportion (15–50%) apparently lacking known CFs or adhesins (Del Canto et al., 2011; Isidean et al., 2011). These strains may be reflecting our inability to reproduce proper conditions for target expression in the laboratory, or recognition inability by antibodies, probes or primers due to mutations generating closely-related variants. Importantly, negative results may reflect the existence of novel adhesins.

Current advances in massive genome sequencing may address these problems. In this work, our aim was to find novel adhesion loci by analyzing the draft genomes of 35 ETEC strains that appeared negative for currently known adhesins, hereon referred to as AN-ETEC strains. We focused particularly on loci encoding pili assembled by the chaperone-usher pathway (CU pili), which includes most of the currently known CFs (Madhavan and Sakellaris, 2015). These structures are composed of two or more structural subunits, the most abundant of which is the major structural subunit (**Figure 1**; Waksman and Hultgren, 2009). The remaining are classified as minor subunits, or tip subunits if they constitute the tip of the pilus (Waksman and Hultgren, 2009). The assembly of the pilus structure requires the action of two types of proteins: the usher and the chaperones. The usher is an outer membrane pore forming protein through which structural subunits are exported (Waksman and Hultgren, 2009). The export process is assisted by one or more chaperones that binds and fold structural subunits in the periplasm (Waksman and Hultgren, 2009). The amino acid sequence of the usher is used to classify CU pili, in a phylogenetic context, into six different families: α, β, γ, κ, π, and σ (Nuccio and Bäumler, 2007). Since the γ CU-family is the most numerous, it has been divided into four subfamilies: γ_1, γ_2, γ_3, and γ_4 (Nuccio and Bäumler, 2007). Fifteen of the 22 ETEC CFs have CU pili assigned to one of these families. Thus, CFA/I, CS1, CS2, CS4, CS5, CS14, CS17, and CS19 belong to the α family; CS12, CS18, and CS20 belong to the γ_2 family; CS3 and CS6 belong to the γ_3 family; and CS13 and CS23 belong to the κ family (Madhavan and Sakellaris, 2015; **Figure 1**). CS7, CS15, and C22 have also been considered CU pili, as they share similarity with known structures from families α (CS7 is similar to CS5) and γ_3 (CS15 and CS22 are similar to SEF14 produced by *Salmonella enterica* serovar Enteritidis; Pichel et al., 2000; Del Canto et al., 2012). However, they have not been assigned to any particular family (Madhavan and Sakellaris, 2015). Consequently, as most of the CFs are CU pili, we expected to find novel CU pili loci, or genetic variants of those already known, in genomes of ETEC strains which lacks CFs.

MATERIALS AND METHODS

Strains

Thirty five ETEC strains were included in the study, each of which were considered negative for known adhesins (AN-ETEC), because no product was obtained in PCR analyses aimed at detecting loci encoding CFA/I, CS1, CS2, CS3, CS4, CS5, CS6, CS7, CS8, CS12, CS13, CS14, CS15, CS17, CS18, CS19, CS20, CS21, CS22, CS23, Tia, TibA, or EtpA. The complete list of primers used in this work is included in **Table S2** (Supporting Information). All AN-ETEC strains had been isolated from children under five years of age with watery diarrhea: 14 of the 35 strains were obtained in Chile and had been previously characterized according to O serogroup and enterotoxin genes (Del Canto et al., 2011). The other 21 isolates were obtained as part of the Global Enterics Multicenter Study (GEMS) in Kenya, Mali, Mozambique, The Gambia, Pakistan, India, and Bangladesh (Kotloff et al., 2013). Enterotoxin gene repertoires were determined using standard procedures as described previously by our research group (Del Canto et al., 2011; Panchalingam et al., 2012). For this group of strains (GEMS' strains), serotyping of the O antigen was carried out by seroagglutination at the Programa de Microbiología y Micología of the Universidad de Chile.

Sequencing

To obtain draft genomic sequences for the 35 AN-ETEC strains, genomic DNA was purified using the Wizard Genomic DNA purification kit (Promega), processed according to the Illumina Paired End Protocol with inserts of 400 bp, and sequenced in a HiSeq 2000 platform at the Institute for Genome Sciences (Baltimore, MD). Reads were analyzed using FastQC v0.10.1 (Babraham Bioinformatics, Babraham Institute, Cambridge, UK. Available: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and assembled using MaSuRCA 2.2.1 or SPAdes 3.1.0 (Bankevich et al., 2012; Zimin et al., 2013). The best of the two assemblies was finally chosen. Best kmer length for assembly was predicted with KmerGenie v.16476 prior to assembly with SPAdes (Chikhi and Medvedev, 2014). Assembly statistics were

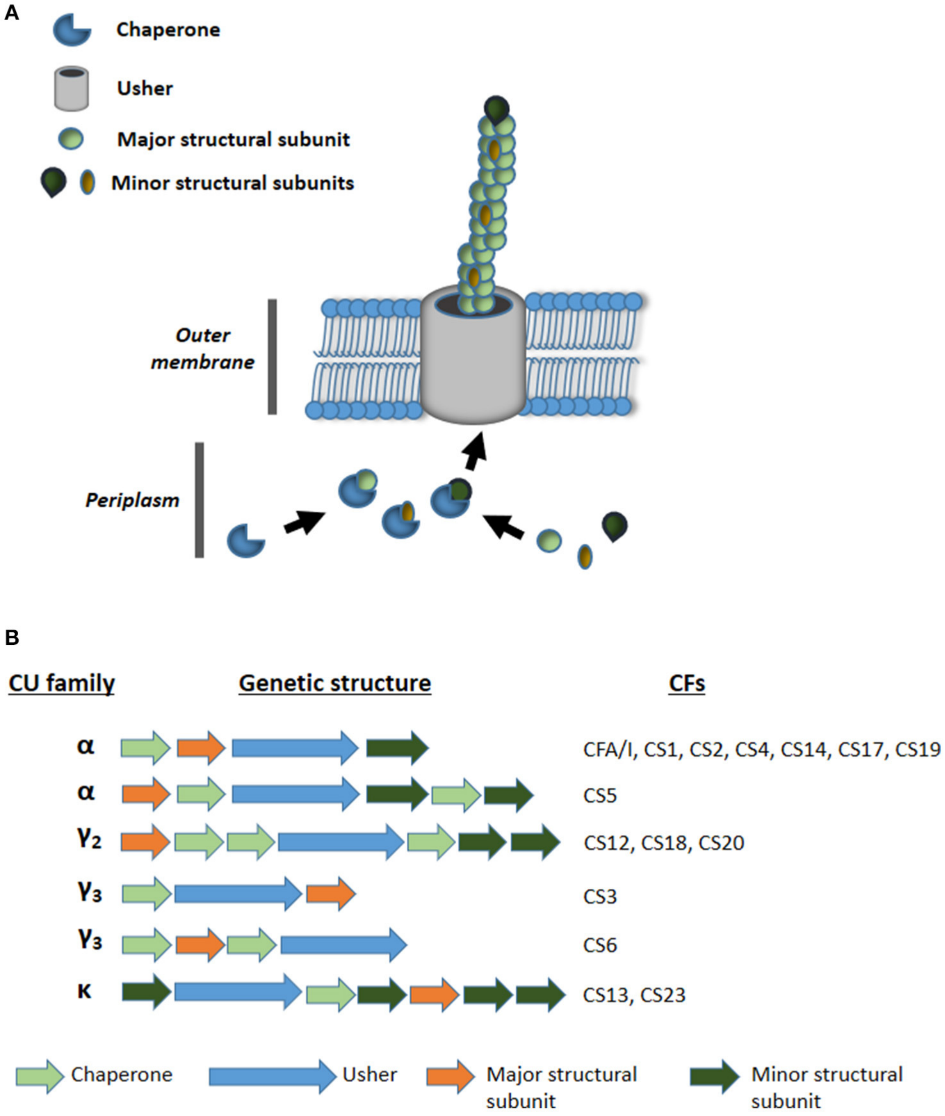


FIGURE 1 | (A) Schematic representation of the assembly of chaperone usher (CU) pili. **(B)** Genetic organization of CU pili loci families α, γ_2, γ_3, and κ encoding known colonization factors (CFs) of ETEC.

obtained with Quast v2.3 and the completeness/contamination report with CheckM (Gurevich et al., 2013; Parks et al., 2015). Sequence annotation was first performed using the Rapid Annotation Subsystem Technology (RAST; Aziz et al., 2008) and then through the NCBI Prokaryotic Genome Annotation Pipeline (National Center for Biotechnology Information, Bethesda MD, USA. Available: http://www.ncbi.nlm.nih.gov/genome/annotation_prok/).

Phylogeny

A phylogenetic tree was inferred based on single nucleotide polymorphisms in draft genomes of our AN-ETEC, by using the CSI phylogeny 1.1, a tool on the Center for Genomic Epidemiology server (Kaas et al., 2014). The following genomic sequences, containing their respective plasmids, were included in the analysis: *E. coli* B, *E. coli* K-12 MG1655, *E. coli* HS, *E. coli* W, *E. coli* ED1a, *E. coli* IAI-1, *E. coli* SE11, *E. coli* SE15 and ETEC prototype strains H10407, E24377A, and B7A. The accession numbers of these sequences are included in **Table S3** of the Supporting Information. The genomic sequence of *E. coli* K-12 MG1655 was set as the reference. A total of 2,481,579 positions were found in all analyzed genomes (53.46% of the reference). The tree was drawn using Mega 6.06 (Tamura et al., 2013). Phylogenetic groups were assigned according to the PCR protocol described by Clermont et al. and complemented with manual analysis of the target genes (Clermont et al., 2013). Sequence types were assigned using the MLST 1.7 of the CGE, according to the scheme proposed by Wirth et al. based on genes *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA* (Wirth et al., 2006; Larsen et al., 2012).

Loci Identification and Analysis

Putative adhesin loci in AN-ETEC genomes were identified first by screening of major CF subunits and then by screening of usher genes using Blast Ring Image Generator (BRIG) v0.95 and/or Large-scale Blast Score Ratio (LS-BSR) (Alikhan et al., 2011; Sahl et al., 2014). In LS-BSR, a blast score ratio (BSR) between 0.00 and 1.00 was obtained for every screened gene in an individual genome. BSR values of 1.00 indicate the maximum sequence identity. For CF screening, genes encoding the following structural subunits were considered: CFA/I, CS1, CS2, CS3, CS4, CS5, CS6, CS7, CS8, CS12, CS13, CS14, CS15, CS17, CS18, CS19, CS20, CS21, CS22, and CS23. For the usher screening, a set of 162 genes of CU pili belonging to the α (16 genes), β (5 genes), γ_1 (25 genes), γ_2 (7 genes), γ_3 (15 genes), γ_4 (40 genes), κ (9 genes), π (25 genes), and σ families (20 genes), were selected based primarily on a previous review (Nuccio and Bäumler, 2007). The complete list of genes used for screening is included in **Table S1** (supporting information). Homolog genes were located within AN-ETEC genomes and their genetic context was analyzed to look for open reading frames (ORFs) encoding putative structural subunits, chaperones, transcriptional regulators, and others that could be part of a CU pili locus. Gene clusters were characterized and compared to known CU pili loci using the NCBI ORF finder (National Center for Biotechnology Information, Bethesda MD, USA. Available: http://www.ncbi.nlm.nih.gov/projects/gorf/), Clustal Omega and/or Unipro UGENE (Sievers et al., 2011; Okonechnikov et al., 2012). In most cases, phylogenetic trees were inferred according to amino acid sequence alignment using Mega 6 (Tamura et al., 2013). Heat maps for screening with LS-BSR or percentage of identity were draw with the Multiple experiment Viewer (MeV) v4.9.0, and loci comparison graphics were drawn with EasyFig v2.1 (Saeed et al., 2006; Sullivan et al., 2011).

Data Availability

The 35 AN-ETEC genomes were partially sequenced as part of an ETEC whole genome shotgun project and deposited in GenBank under the accession BioProject PRJNA287625. In addition, a file containing the amino acid sequences of the major structural subunits was included as Supplementary Material (Supplementary **Data Sheet 1**). Chilean bacterial isolates may be requested by contacting corresponding authors and GEMS isolates may be requested at http://www.medschool.umaryland.edu/GEMS/GEMS-Data--Specimen-Requests/.

RESULTS

Thirty-five AN-ETEC strains genomes were sequenced in order to identify putative adhesin loci. These strains belonged to 13 different somatic serogroups (O serogroup) of which 10 were non-typeable (ONT), and 82% (29/35) were positive for LT only (lacking ST). Sequencing and assembly statistics are shown in **Table 1**. Sequencing coverage ranged between 75X and 353X, and the average length of the assembled draft genomes was 5,126,274 bp, which was close to the expected size for a pathogenic *E. coli* strain (Lukjancenko et al., 2010). Estimates of completeness and contamination for our draft genomes, assessed using a set of *Enterobacteriaceae*-specific markers, indicated that the vast majority of them were nearly complete (≥95%) and had low levels of contamination (≤5%). Only for two strains, ETEC 1241a and ETEC 4155a, results suggest medium contamination levels (6.04 and 5.04%, respectively). Phylogenetic characterization of our AN-ETEC strains, according to their genomes by SNP calling and *in silico* multiple locus sequence typing (MLST), was generally consistent with phylogenetic groups and serogroups, rather than with the geographical origin of the strains. Three main groups of strains can be identified in the tree shown in **Figure 2**. One group included 10 AN-ETEC strains primarily isolated in Chile, belonging to phylogroup A and sequence types ST-100, ST-750, and ST-165. The second group included 10 AN-ETEC strains belonging also to the phylogroup A, but associated with eight different sequence types. Non-pathogenic strains *E. coli* B, *E. coli* K-12 MG1655, *E. coli* HS, and the prototype ETEC H10407, all known members of the phylogroup A (Wurpel et al., 2013), were located in this group. The third group contained 14 strains belonging to the phylogroup B_1. This group was associated with eight different sequence types, none of which are present in phylogroup A AN-ETEC, including known non-pathogenic *E. coli* strains, *E. coli* W, *E. coli* SE11, *E. coli* IAI1, and the prototypic ETEC strains ETEC E24377A and ETEC B7A (Wurpel et al., 2013). Only one of our AN-ETEC (ETEC 200617) belonged to the phylogroup D. Overall, the AN-ETEC collection analyzed in this work is genetically diverse belonging mainly to phylogroups A and B_1. No widely distributed or predominant serogroup or sequence type was identified.

CFs-Bioinformatical Screening

Strains in this study had been previously characterized by PCR and all of them had tested negative for known ETEC CFs genes. We used the same genes used in the PCR screening to probe the draft genomes using LS-BSR with the tblastn algorithm (Sahl et al., 2014). As seen in **Figure 3**, CF genes were not detected in non-pathogenic *E. coli* strains, with blast-score ratio (BSR) values ranging between 0.00 and 0.32. In contrast, they were correctly detected in ETEC prototype strains H10407 (CFA/I), ETEC E24377A (CS1 and CS3), and ETEC B7A (CS6) with BSR values ranging between 0.97 and 1.00. Although most of these AN-ETEC genomes were confirmed negative for known CF genes by this second screening method, the blast-score ratio (BSR) values obtained in some cases (between 0.57 and 0.79) for genes encoding CS8/CS21 (ETEC 401909), CS20 (ETEC 8255a, ETEC 8350a-1, ETEC200144, ETEC200144, ETEC 200617, ETEC 702251 and ETEC 505077), and CS23 (ETEC 100664, ETEC 402594, ETEC 702251 and ETEC 703098), suggest the presence of homologs to those genes.

CS8 and CS21 are homologous structures belonging to the type IV pili family (Madhavan and Sakellaris, 2015). Interestingly, a gene homologous to *cofA* and *lngA*, encoding CS8 and CS21 pilins, respectively, was localized in the ETEC 401909 strain (Kenya, O9, phylogroup C, ST-410, LT), the only AN-ETEC strain with relatively high BSR values (0.79 for CS8 and 0.75 for CS21). Analysis of the PCR primers' target sequences for *cofA* and *lngA* showed discrepancies that may explain negative PCR

TABLE 1 | Summary of draft genomic sequences of ETEC strains obtained in this work.

Strain	Origin	Coverage*	Assembler	Total length (bp)	Contigs	N50	% GC	% Completeness†	% Contamination†	GenBank accession code
3693	Chile	226	SPAdes	5,001,931	251	75,875	50.6	99.96	0.40	LGLX00000000
12684a	Chile	182	MaSuRCA	4,887,648	259	84,003	50.9	95.33	2.56	LGLY00000000
15617a	Chile	149	MaSuRCA	4,979,396	149	133,333	50.8	96.63	2.50	LGLZ00000000
963a-1	Chile	144	MaSuRCA	5,000,982	193	108,211	50.5	92.44	3.20	LGMA00000000
1030c-1	Chile	157	MaSuRCA	5,318,135	203	86,519	50.6	99.65	2.80	LGMB00000000
1241a	Chile	179	MaSuRCA	5,498,099	250	102,427	50.8	100	6.04	LGMC00000000
4155a	Chile	175	MaSuRCA	5,052,165	171	104,482	50.6	97.58	5.04	LGMD00000000
8255a	Chile	213	MaSuRCA	5,029,983	289	81,830	50.6	99.65	2.30	LGME00000000
8350a	Chile	170	MaSuRCA	5,024,877	136	138,088	50.7	99.34	0.29	LGMF00000000
9312a	Chile	220	SPAdes	4,940,375	158	150,786	50.6	100	0.08	LGMG00000000
9343a	Chile	120	MaSuRCA	5,049,524	146	151,450	50.7	100	1.22	LGMH00000000
9571a	Chile	207	MaSuRCA	5,119,006	233	137,783	50.7	99.96	2.07	LGMI00000000
9788a	Chile	172	MaSuRCA	4,954,530	130	134,873	50.7	99.73	0.69	LGMJ00000000
10380a	Chile	187	MaSuRCA	4,825,041	147	117,204	50.6	99.96	1.62	LGMK00000000
700377	Pakistan	187	MaSuRCA	5,509,751	282	89,754	50.4	98.06	2.74	LGML00000000
300659	Mozambique	146	MaSuRCA	5,193,583	176	116,939	50.6	99.96	2.37	LGMM00000000
200144	Mali	127	MaSuRCA	5,096,812	211	75,533	50.8	99.96	1.70	LGMN00000000
200617	Mali	168	MaSuRCA	5,474,100	183	177,338	50.3	99.66	1.28	LGMO00000000
503210	India	75	MaSuRCA	5,170,803	119	134,854	50.7	99.96	2.25	LGMP00000000
401909	Kenya	75	MaSuRCA	5,018,235	110	133,005	50.5	99.96	1.19	LGMQ00000000
401900	Kenya	343	SPAdes	4,737,288	139	133,706	50.6	99.96	0.11	LGMR00000000
100664	The Gambia	103	MaSuRCA	5,332,273	201	80,008	50.4	99.96	4.66	LGMS00000000
302025	Mozambique	127	MaSuRCA	5,152,571	172	114,889	50.6	99.96	2.36	LGMT00000000
602687	Bangladesh	109	MaSuRCA	5,029,988	242	67,774	50.8	99.39	2.42	LGMU00000000
402594	Kenya	166	MaSuRCA	5,280,101	173	94,074	50.5	99.96	2.80	LGMV00000000
702251	Pakistan	145	MaSuRCA	5,323,290	334	55,624	50.4	99.34	3.56	LGMW00000000
702332	Pakistan	102	MaSuRCA	5,013,771	167	88,063	50.8	99.96	2.49	LGMX00000000
703098	Pakistan	353	SPAdes	5,356,213	300	168,561	50.6	99.96	0.15	LGMY00000000
603388	Bangladesh	126	MaSuRCA	5,166,599	160	162,966	50.6	99.96	2.85	LGMZ00000000
203518	Mali	165	MaSuRCA	5,036,223	137	130,574	50.8	99.96	1.50	LGNA00000000
403418	Kenya	169	MaSuRCA	5,069,930	135	119,990	50.8	99.96	1.84	LGNB00000000
703200	Pakistan	78	MaSuRCA	5,249,204	128	124,962	50.4	96.26	3.72	LGNC00000000
603936	Bangladesh	150	MaSuRCA	5,224,195	138	120,925	50.7	99.96	1.88	LGND00000000
505077	India	131	MaSuRCA	5,314,101	194	125,041	50.6	98.73	4.55	LGNE00000000
403885	Kenya	131	MaSuRCA	4,988,884	134	121,380	50.9	100	0.47	LGNF00000000

*Coverage estimated according to the ETEC H10407 genome length, including plasmids.
†Estimated by presence of lineage markers using CheckM.

results (**Figure S1A** in Supporting Information). The full-length gene nucleotide sequence shared 77% identity with *cofA* and 60% identity with *lngA*, but was 100% identical with the recently discovered CS8b variant, which has yet to be functionally characterized (Njoroge et al., 2015). The alignment of the derived amino acid sequences allowed us to probe this identity (**Figure S1B**). Therefore, even though the ETEC 401909 strain was negative for detection of *cofA* and *lngA*, it harbored a recently discovered and related variant.

CU Pili

Most of the CFs are CU pili. By screening of the CF structural subunit genes that were targets for PCR detection in AN-ETEC strains, we were able to identify sequences with high similarity (**Figure 3**). In order to complement this analysis and look for additional potential loci homologous to known CU pili loci, we bioinformatically screened a set of 162 usher genes from AN-ETEC genomes, using the blast- ring image generator (BRIG) with the blastn option and LS-BSR with tblastn (Alikhan et al., 2011; Sahl et al., 2014). Usher genes were chosen because they are present only in a single copy within the CU loci. This analysis included the genomes of 13 non-pathogenic *E. coli* strains and three ETEC prototype strains. As shown in **Figure 4**, the results derived from both approaches were similar. As previously reported, there is a set of core CU pili loci in *E. coli* including *mat* (α family), *fim* (γ_1), *yde* (γ_1), *yad* (γ_4), *yeh* (γ_4), *yfc* (π), and *ybg*

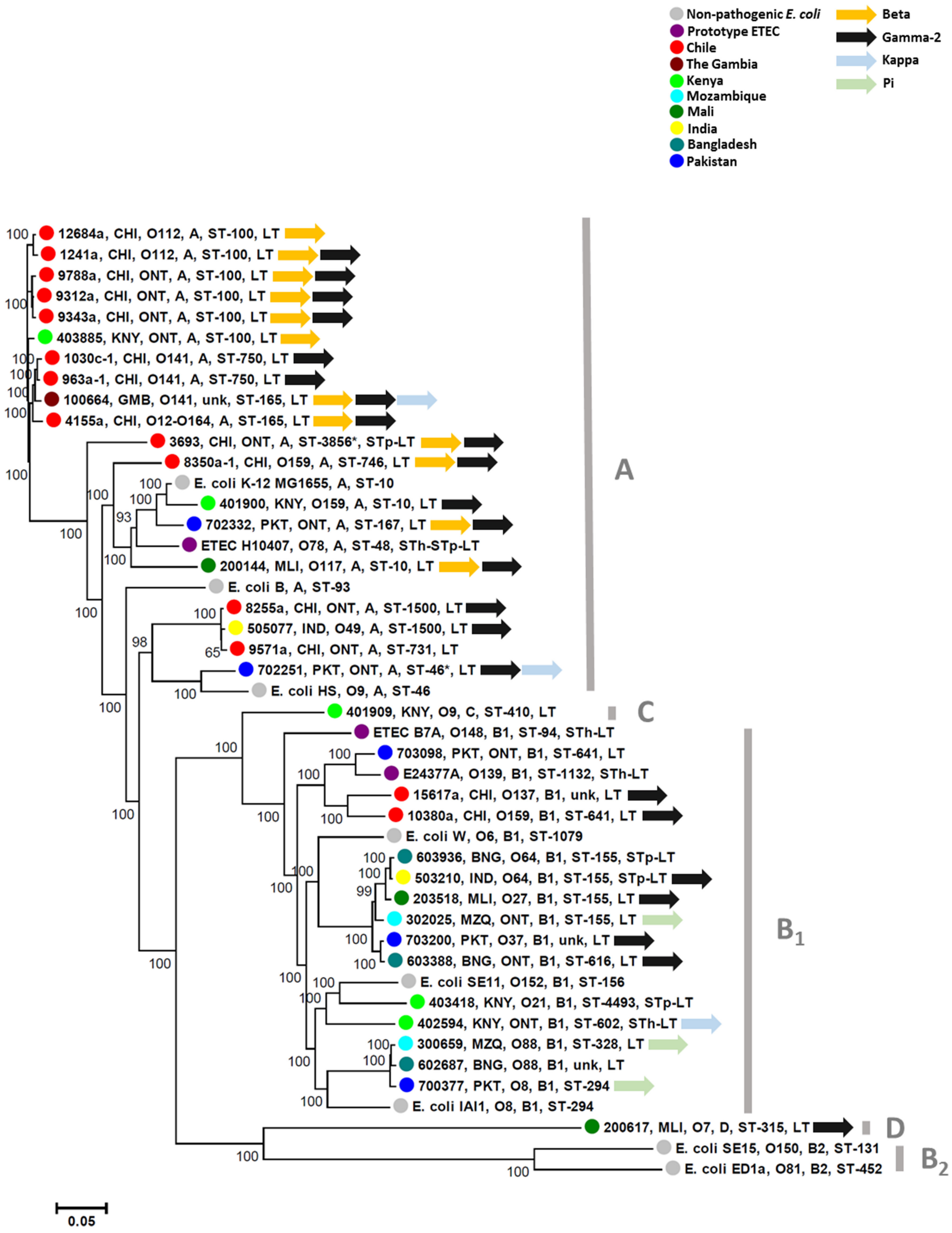

Non-pathogenic E. coli
Prototype ETEC
Chile
The Gambia
Kenya
Mozambique
Mali
India
Bangladesh
Pakistan
Beta
Gamma-2
Kappa
Pi
12684a, CHI, O112, A, ST-100, LT
1241a, CHI, O112, A, ST-100, LT
9788a, CHI, ONT, A, ST-100, LT
9312a, CHI, ONT, A, ST-100, LT
9343a, CHI, ONT, A, ST-100, LT
403885, KNY, ONT, A, ST-100, LT
1030c-1, CHI, O141, A, ST-750, LT
963a-1, CHI, O141, A, ST-750, LT
100664, GMB, O141, unk, ST-165, LT
4155a, CHI, O12-O164, A, ST-165, LT
3693, CHI, ONT, A, ST-3856*, STp-LT
8350a-1, CHI, O159, A, ST-746, LT
E. coli K-12 MG1655, A, ST-10
401900, KNY, O159, A, ST-10, LT
702332, PKT, ONT, A, ST-167, LT
ETEC H10407, O78, A, ST-48, STh-STp-LT
200144, MLI, O117, A, ST-10, LT
E. coli B, A, ST-93
8255a, CHI, ONT, A, ST-1500, LT
505077, IND, O49, A, ST-1500, LT
9571a, CHI, ONT, A, ST-731, LT
702251, PKT, ONT, A, ST-46*, LT
E. coli HS, O9, A, ST-46
401909, KNY, O9, C, ST-410, LT
ETEC B7A, O148, B1, ST-94, STh-LT
703098, PKT, ONT, B1, ST-641, LT
E24377A, O139, B1, ST-1132, STh-LT
15617a, CHI, O137, B1, unk, LT
10380a, CHI, O159, B1, ST-641, LT
E. coli W, O6, B1, ST-1079
603936, BNG, O64, B1, ST-155, STp-LT
503210, IND, O64, B1, ST-155, STp-LT
203518, MLI, O27, B1, ST-155, LT
302025, MZQ, ONT, B1, ST-155, LT
703200, PKT, O37, B1, unk, LT
603388, BNG, ONT, B1, ST-616, LT
E. coli SE11, O152, B1, ST-156
403418, KNY, O21, B1, ST-4493, STp-LT
402594, KNY, ONT, B1, ST-602, STh-LT
300659, MZQ, O88, B1, ST-328, LT
602687, BNG, O88, B1, unk, LT
700377, PKT, O8, B1, ST-294
E. coli IAI1, O8, B1, ST-294
200617, MLI, O7, D, ST-315, LT
E. coli SE15, O150, B2, ST-131
E. coli ED1a, O81, B2, ST-452
A
C
B_1
D
B_2
0.05

FIGURE 2 | Phylogenetic diversity of ETEC strains included in this study. Tree depicting genetic relationships among AN-ETEC strains based on a single nucleotide polymorphism-calling procedure (CSI phylogeny). Font colors indicate different geographical origins (countries). In each node, the indicated features are as follows: strain code, country of origin, O serogroup, phylogroup, sequence type, and enterotoxins. The following non-pathogenic strains (in black fonts) and ETEC prototype strains (in purple fonts) were included in the analysis: *E. coli* B (REL606), *E. coli* W (Crooks), *E. coli* K-12 MG1655, *E. coli* HS, *E. coli* ED1a, *E. coli* IAI1, *E. coli* SE11, *E. coli* SE15, ETEC H10407, ETEC E24377A, and ETEC B7A. Asterisks in ETEC 3693 and ETEC 702251 indicate sequence types suggested by the MLST 1.7 tool when non-perfect matching occurred. Unk, unknown sequence type.

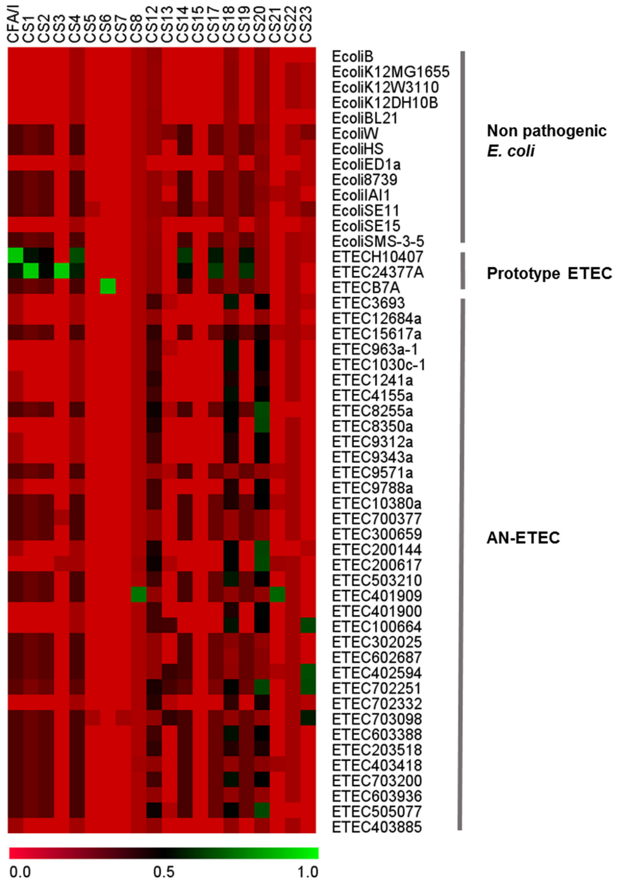


FIGURE 3 | Bioinformatical screening of colonization factors. Heat map derived from the screening of CF-structural subunit genes with large scale-blast score ratio (LS-BSR). Genomes of non-pathogenic and ETEC prototype strains were included in the analysis.

(π) (Wurpel et al., 2013). Usher genes from all of these loci were detected in both non-pathogenic and pathogenic ETEC genomes. Usher genes from loci encoding known CFs of ETEC prototype strains were detected, namely CFA/I of ETEC H10407 and CS1 of ETEC E24377A in the α family, along with CS3 of ETEC E24377A and CS6 of ETEC B7A in the γ_3 family. Consistent with CF gene screening, homologous sequences encoding ushers from the γ_2 and κ families were found in several AN-ETEC strains but were not detected in the majority of the non-pathogenic strains. Similar results were observed for usher genes from the β family and from some representatives of the π family. No homologs of usher genes from the σ family were detected in our set of AN-ETEC genomes. Our results suggest that AN-ETEC strains may contain CU-pili loci encoding structures from the β, γ_2, κ, and π families.

Kappa CU Pili Family

According to our CF gene screening, genes homologous to *aalE*, which encodes the CS23 major subunit, would be harbored by four AN-ETEC strains (and were not restricted to any particular phylogroup or ST): ETEC 100664, ETEC 402594, ETEC 702251, and ETEC 703098. This observation was consistent with the fact that sequences homologous to usher genes belonging to the kappa family were detected in these strains (BSR values between 0.41 and 0.99, **Figure 4B**). However, the *aalE* primer did not match perfectly in any of these genomes, which explains the negative PCR results (**Figure 5A**). Based on the organization of the genetic cluster encoding CS23 and CS13 (Nuccio and Bäumler, 2007; Madhavan and Sakellaris, 2015), complete loci were identified in three of the four mentioned AN-ETEC strains (ETEC 100664, ETEC 402594, and ETEC 702251). This includes six genes encoding structural subunits, one for a chaperone, and one for the usher (Nuccio and Bäumler, 2007; Madhavan and Sakellaris, 2015). In the case of the ETEC 703098 strain, the sequencing did not cover the complete cluster and only five structural subunits were found: the chaperone and the 3′ half portion of the usher gene. Alignment of the amino acid sequences of the putative major structural subunits (PMSs) indicated that these identified loci would encode pili more closely-related to CS23 than to CS13 (**Figure 5B**). Identity percentages between AalE and the novel PMSs were 68% for ETEC 402594, 67% for ETEC 100664 and ETEC 702251, and 59% for ETEC 703098. As κ-CU pili loci are also present in the genomes of non-pathogenic *E. coli* W and *E. coli* SE11, we included these in the analysis. However, the κ-PMS sequences derived from both strains are different from those found in AN-ETEC strains, sharing only 26 and 25% amino acid identity with AalE (**Figure 5B**). The regions of identity among all the κ-PMSs resided mainly in the first 30 amino acids, which corresponds with the predicted AalE signal sequence for secretion to the periplasm (Del Canto et al., 2012).

Analysis of the κ-CU pili loci of AN-ETEC strains showed that they share the same genetic organization (**Figure 5C**). A gene encoding a putative transcriptional regulator, similar to that harbored in the CS23 locus, was found in the three AN-ETEC strains having a complete κ-CU pili locus. In addition, a putative seventh structural subunit gene was found in the ETEC 100664 and ETEC 702251 κ-CU pili loci, in which the contig length allowed for recognition of downstream sequences. This was located downstream a gene encoding a putative phosphodiesterase (EAL domain-containing protein). Alignment using tblastx indicated that κ-CU pili loci carried by ETEC 100664 and ETEC 702251 are nearly identical, and differ from loci harbored by ETEC 402594 and the *aal* locus, mainly in the PMS sequence. A higher level of dissimilarity for all ORFs was observed between the κ-CU pili loci of AN-ETEC strains, the *csh* locus (CS13), and κ-CU pili loci of *E. coli* W and *E. coli* SE11.

Screening of usher genes also allowed us to identify a locus similar to the κ-CU pili loci, encoding fimbriae AF/R1 of rabbit enteric pathogen *E. coli* RDEC-1 and F18 of porcine ETEC in the AN-ETEC 602687 strain. However, as sequencing did not allow for assembly of a complete locus, we did not analyze this cluster.

Gamma-2 CU Pili Family

Genomes of six AN-ETEC strains showed high BSR scores for a gene encoding the major structural subunit of CS20, suggesting the presence of γ_2-CU pili loci in AN-ETEC strains (**Figure 3**).

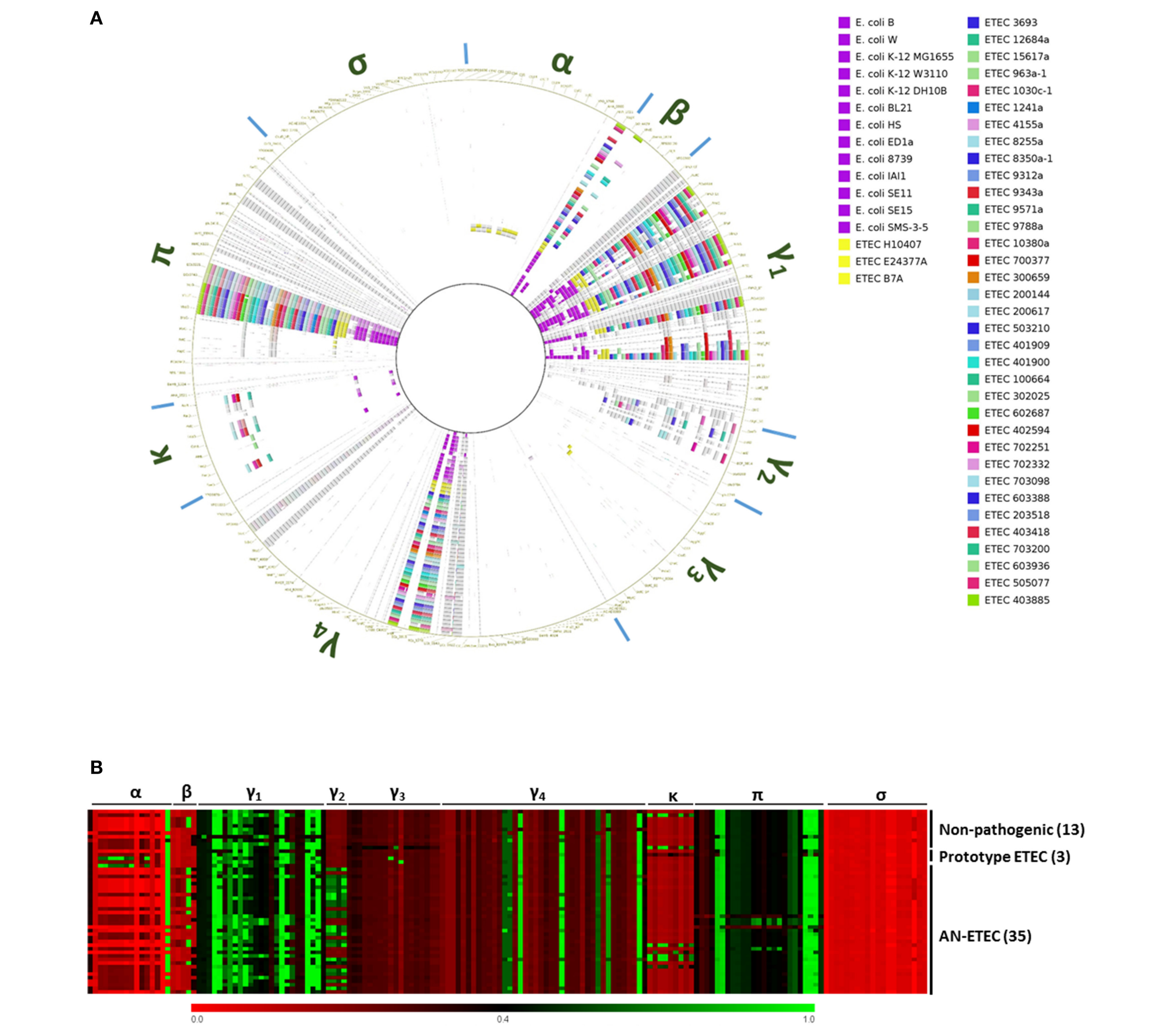


FIGURE 4 | Bioinformatical screening for chaperone-usher (CU) pili loci among AN-ETEC strains. (A) Screening of usher genes using blastn in blast ring image generator (BRIG). Cut-off values were 70% for identity percentage and 0.01 for the blast *E*-value. Greek letters around the ring indicate columns corresponding to the different CU-pili families. **(B)** Heat map derived from the screening of the same set of usher genes using tblastn in LS-BSR. Columns corresponding to the different CU-pili families are indicated at the top of the map. In **(A,B)** genomes of non-pathogenic *E. coli* strains and ETEC prototype strains were included.

After complementing this information with the results from our usher genes screening, γ_2-CU pili loci were identified in 23 AN-ETEC genomes (BSR values ranging between 0.46 and 1.00, **Figure 4B**). According to the genetic organization of loci encoding γ_2-CFs CS12, CS18, and CS20, the loci identified in AN-ETEC strains do have the minimum number of genes required to direct synthesis of a functional structure, including three encoding structural subunits, three for chaperones, and one for an usher (Nuccio and Bäumler, 2007; Madhavan and Sakellaris, 2015). However, alignment of amino acid sequences of the PMSs proved that none were identical to the currently known major structural subunits CswA (CS12), FotA (CS18), or CsnA (CS20) and that most of these were closely related to the last two (**Figure 6A**; Nuccio and Bäumler, 2007; Madhavan and Sakellaris, 2015). Overall, there were six groups derived from the alignment, containing between two and five 100% identical PMS sequences. Putative γ_2-PMS from two of these groups (PMS harbored by ETEC 200617, ETEC 505707, ETEC 8255a and ETEC 200144, ETEC 702251) plus that harbored by ETEC 8350a-1, shared about 76% identity with CsnA. Five of these six strains belong to the phylogroup A and one to the phylogroup D, representing four different serogroups and five sequence types.

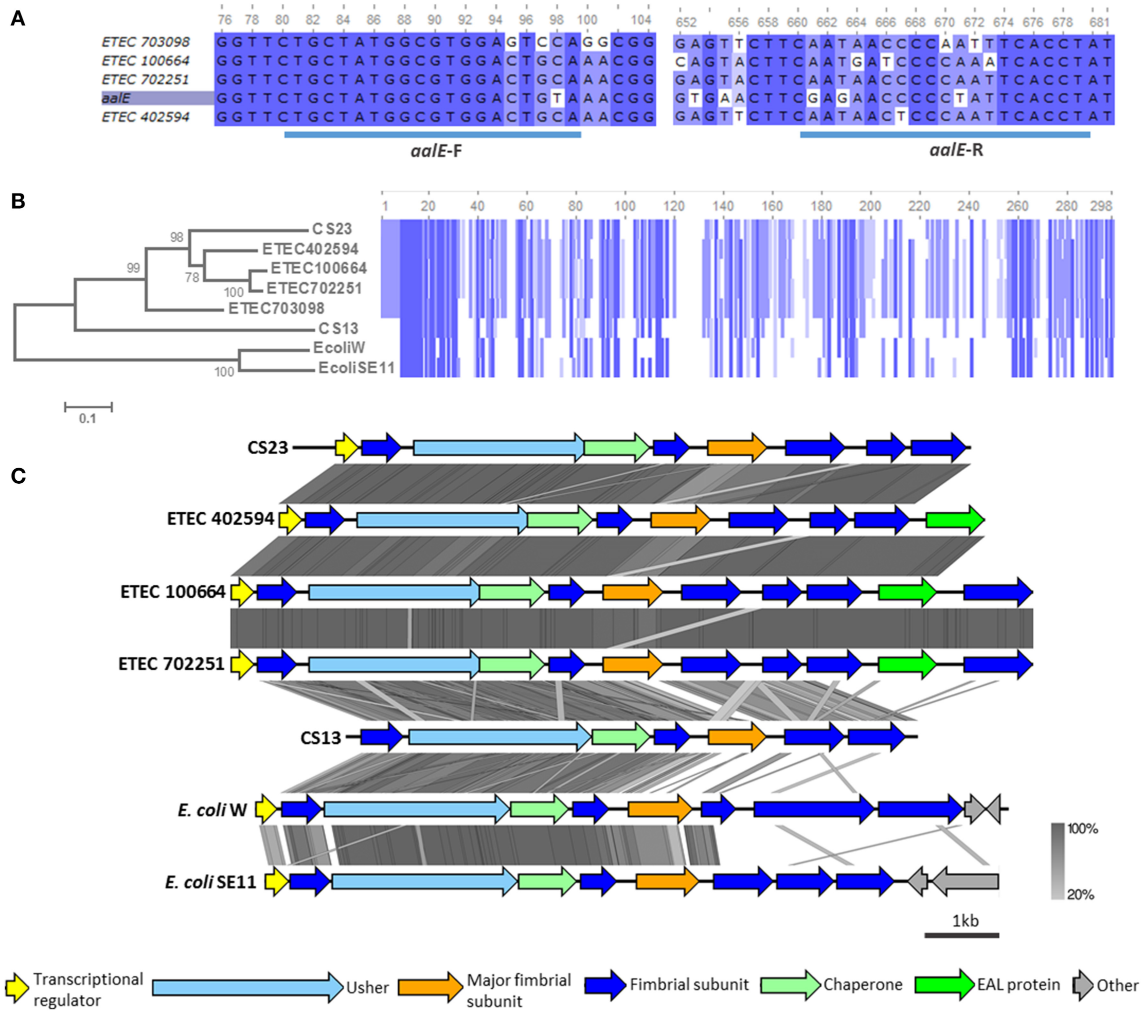


FIGURE 5 | Analysis of κ-CU pili loci found in AN-ETEC strains. (A) Alignment of the gene encoding CS23 major subunit (*aalE*) with homologs found in AN-ETEC strains. The blue bars at the bottom indicate the primer's target sequences of forward and reverse primers. **(B)** Alignment of the amino acid sequences of AalE and the homologs found in AN-ETEC strains. Putative major subunits of κ-CU pili of non-pathogenic strains *E. coli* W and *E. coli* SE11 were included. Blue blocks indicate identical regions and the color fades as the identity percentage decreases. The tree was inferred from alignment using the neighbor joining method with 1000 bootstrap replicates. **(C)** Alignment of predicted κ-CU pili loci found in AN-ETEC strains using tblastx in Easyfig. Loci encoding CS23 and CS13, such as those κ-CU pili loci found in *E. coli* W and *E. coli* SE11, were included in the analysis.

Eight AN-ETEC strains (the group of ETEC 4155a, ETEC 603388, ETEC 1030c-1, ETEC 963a-1, ETEC 703200, the group of ETEC 3693, ETEC 503210, plus ETEC 100664) harbor γ_2-PMS with 58–59% identity to FotA (**Figure 6A**). This group includes five phylogroup A strains and three phylogroup B_1 strains. Other five strains (ETEC 10380a and the group of ETEC 9312a, ETEC 9343a, ETEC 9788a, ETEC 401900) harbor γ_2-PMS that share 59% identity with CsnA. Four of these strains belong to phylogroup A and three are closely related to ST-100 strains isolated in Chile. Within the other four AN-ETEC strains carrying γ_2-CU pili loci, there are two identical PMS sharing 51% identity with CsnA (ETEC 1517a and ETEC 203518) and two different sequences with 50 and 49% identity to CsnA (ETEC 1241a and ETEC 702332). Alignment revealed a block of eight amino acids conserved across CS20 and all the γ_2-PMS of the AN-ETEC strains between positions 50 and 57 (**Figure 6A**). This region corresponds to the residues 18–25 of the mature CsnA protein (Valvatne et al., 1996). Other 33 conserved residues among all the γ_2-PMS and the known γ_2-CFs major subunits were found distributed along the whole extension of the sequences and including single residues, as stretches ranging between two and four amino acids (**Figure 6A**).

Analysis of the genetic structure and alignment of the complete γ_2-CU pili loci using tblastx showed that the structure is maintained across AN-ETEC strains and those identity-percentages are consistent with groups described based on γ_2-PMS sequences (**Figure 6B**). Differences between CS20 loci and those γ_2-CU pili loci included in the first group of six

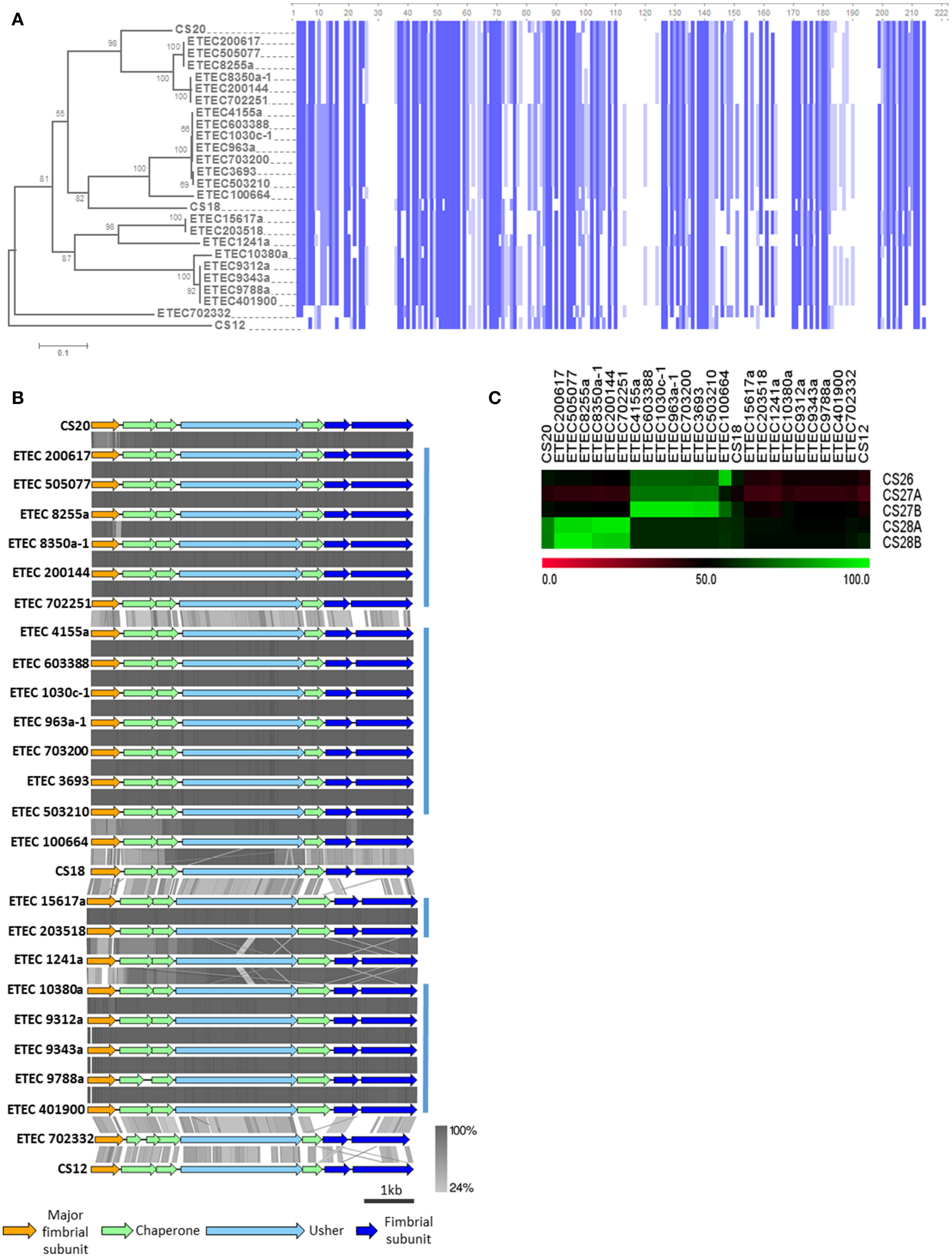


FIGURE 6 | Analysis of γ_2-CU pili loci found in AN-ETEC strains. (A) Alignment of the amino acid sequences of CS12, CS18, and CS20 major structural subunits, and the homologs found in AN-ETEC strains. Blue blocks indicate identical regions and the color fades as the identity percentage decreases. The tree was inferred from alignment using the neighbor joining method with 1000 bootstrap replicates. **(B)** Alignment of the predicted γ_2-CU pili loci found in AN-ETEC strains using tblastx in Easyfig. Loci encoding CS12, CS18, and CS20 were included. Blue bars at the right indicate groups of loci sharing high identity percentages. **(C)** Identity based-heap map derived from the alignment of nucleotide sequences of CS26, CS27a, CS27b, CS28a, and CS28b with γ_2-CU pili putative major subunit genes found in AN-ETEC strains. Colors represent percent of identity.

AN-ETEC strains resided mainly in the PMS, while, in the rest of the γ_2-CU pili, it also depends on the putative chaperones, minor structural subunits and even usher genes.

We aligned sequences with those of five, previously described, partial γ_2-PMS gene sequences: CS26, CS27A, CS27B, CS28A, and CS28B (Nada et al., 2011), using blastn, and found both high similarities and identities. The gamma-2 PMS gene found in ETEC 100664 shared 91% identity with CS26 (**Figure 6C**). CS27A did not share high identity ($\geq$90%) with any gene sequences in our study. In contrast, a group of seven AN-ETEC strains harbored γ_2-PMS genes containing sequences highly similar (94–96%) to CS27B. This group also includes strains whose γ_2-PMS shared 58–59% identity with FotA. Three AN-ETEC bear γ_2-PMS genes sharing high identity with CS28A (96% identity), as was other three AN-ETEC bear γ_2-PMS genes almost identical to CS28B with (97% identity). Together, these results suggest that γ_2-CU pili, other than CS12, CS18, and CS20, might be part of the ETEC adhesin repertoire.

Pi CU Family

None of the known ETEC CFs exhibit π-CU pili, but in our screening of usher genes, we identified three strains from the B_1 phylogroup that were positive for homologs of the π-CU pili family (BSR values ranging between 0.70 and 0.82, **Figure 4B**), namely *sfpC*, *pixC*, *prfC*, and *papC*. These usher genes represent loci that differ from the π-CU pili loci *yfc*, *ybg*, and *yqi*, which were found to be widely distributed among *E. coli* strains (Wurpel et al., 2013). Two of these strains, ETEC 700377 and ETEC 300659, were closely related and were classified as ST-328, while the third, ETEC 302025, was classified as ST-155. Localization of the loci in these three genomes allowed for identification of clusters composed of seven ORFs that encoded four structural subunits, two chaperones and the usher. Alignment of the amino acid sequences derived from the PMS gene using known π-major structural subunits PapA, PrfA, PixA, and SpfA revealed a higher degree of identity with SpfA (67% identity with π-PMS of ETEC 700377 and ETEC 300659, and 65% with π-PMS of ETEC 302025; **Figure 7A**). Four stretches of four amino acids in the N-terminal region of the predicted mature proteins (Brunder et al., 2001), and one of seven amino acids in the C-terminal region found to be conserved across all the sequences included in the analysis (**Figure 7A**). Aligning the genetic clusters using tblastx showed similarities to known loci residing mainly in genes encoding the PMS, the adjacent gene encoding a putative structural subunit gene, the usher, and the two chaperones, although differing in the downstream putative structural subunit genes. In contrast, among the AN-ETEC π-CU pili loci (ETEC 700377/ETEC 300659 v/s ETEC 302025), dissimilarity resided mainly in the PMS and the adjacent putative subunit gene (**Figure 7B**).

A recently described hybrid enteropathogenic *E. coli* (EPEC)/ETEC strain harbors a π-CU pili locus, identical to the locus harbored by ETEC 700377 and ETEC 300659 (Dutta et al., 2015). Therefore, we computationally screened a set of EPEC genes in the three AN-ETEC strains in which we found π-CU pili loci; these genes were typically contained in the locus of enterocyte effacement (LEE), a pathogenicity island carried by both EPEC and enterohemorrhagic *E. coli* (Croxen et al., 2013). According to LS-BSR screening using tblastn, almost all of the EPEC genes are present in ETEC 700377 and ETEC 300659, except for *bfp* (**Figure 7C**). In contrast, none of the EPEC genes were found in ETEC 302025, suggesting that the presence of π-CU pili loci is not restricted to EPEC/ETEC hybrids.

Beta CU Family

To our knowledge, CU pili belonging to the β family have never been characterized, and the denomination as an actual pili family has been only been supported by sequence data (Nuccio and Bäumler, 2007). Although the presence of a β-CU pili locus in *E. coli* K-12 was reported, it is disrupted by an insertion sequence (IS), which may be why it has not received more attention (Korea et al., 2010). Our usher gene screening shows that 15 AN-ETEC genomes, in addition to the three *E. coli* K-12 substrains, also carry β-CU pili loci (BSR values ranging between 0.90 and 1.00, **Figure 4B**). All these strains belong to the phylogroup A. Based on the structure of the locus in *E. coli* K-12 substrain MG1655, 13 AN-ETEC genomes carry a complete cluster, lacking the IS and including five genes: *gltF*, a putative structural subunit; *yhcA*, a putative chaperone; *yhcD*, the putative usher; *yhcE*, whose function is unclear; and *yhcF*, another putative structural subunit (Nuccio and Bäumler, 2007; Korea et al., 2010). In one of the AN-ETEC strains (ETEC 401900) a complete locus was found, but a premature stop codon disrupts the *yhcD* gene. For this reason, this strain was not considered in further analyses.

All of the β-CU pili loci identified share the same organization, with the exception that the insertion sequence disrupting *yhcE* in *E. coli* K-12 is absent in all AN-ETEC genomes (**Figure 8A**). According to the alignment performed with tblastx, the *gltF*, *yhcA*, *yhcD*, and *yhcF* sequences are highly conserved among *E. coli* K-12 and AN-ETEC strains (**Figure 8A**). Only in the ETEC 963a-1 strain, in which sequencing did not allow for the identification of a complete β-CU pili locus, a longer form of *yhcE* was found. This variant likely encodes a protein with additional 57 amino acids in the C-term, unlike the form found in the other AN-ETEC strains. The variability among the identified β-CU pili loci can be almost exclusively attributed to *yhcF*, which was found in three different forms. One of them, of 717 bp, was found in nine AN-ETEC strains and corresponds to the form present in *E. coli* K-12. Two other shorter forms, of 633 and 405 bp, were identified in two and three AN-ETEC genomes, respectively (**Figure 8A**). Alignment of the predicted protein products is consistent with such an observation (**Figure 8B**). The three forms have an identical block of 118 amino acids and the divergence is determined by the length of the C-terminal regions.

In summary, at least one pili locus, different from those that seem to be widely distributed among *E. coli*, was identified in 33 of the 35 AN-ETEC strains included in this study. No loci were identified in strains ETEC 9571a (ONT, phylogroup A, ST-731, LT) and ETEC 603936 (O64, phylogroup B_1, ST-155, STp-LT). We identified 10 different CU pili loci among the 35 AN-ETEC strains analyzed, which, according to the number of ORFs they contain, could encode functional pili structures. A list of these loci, along with their organization and positive strains, is provided in **Table 2**. None of the AN-ETEC strains

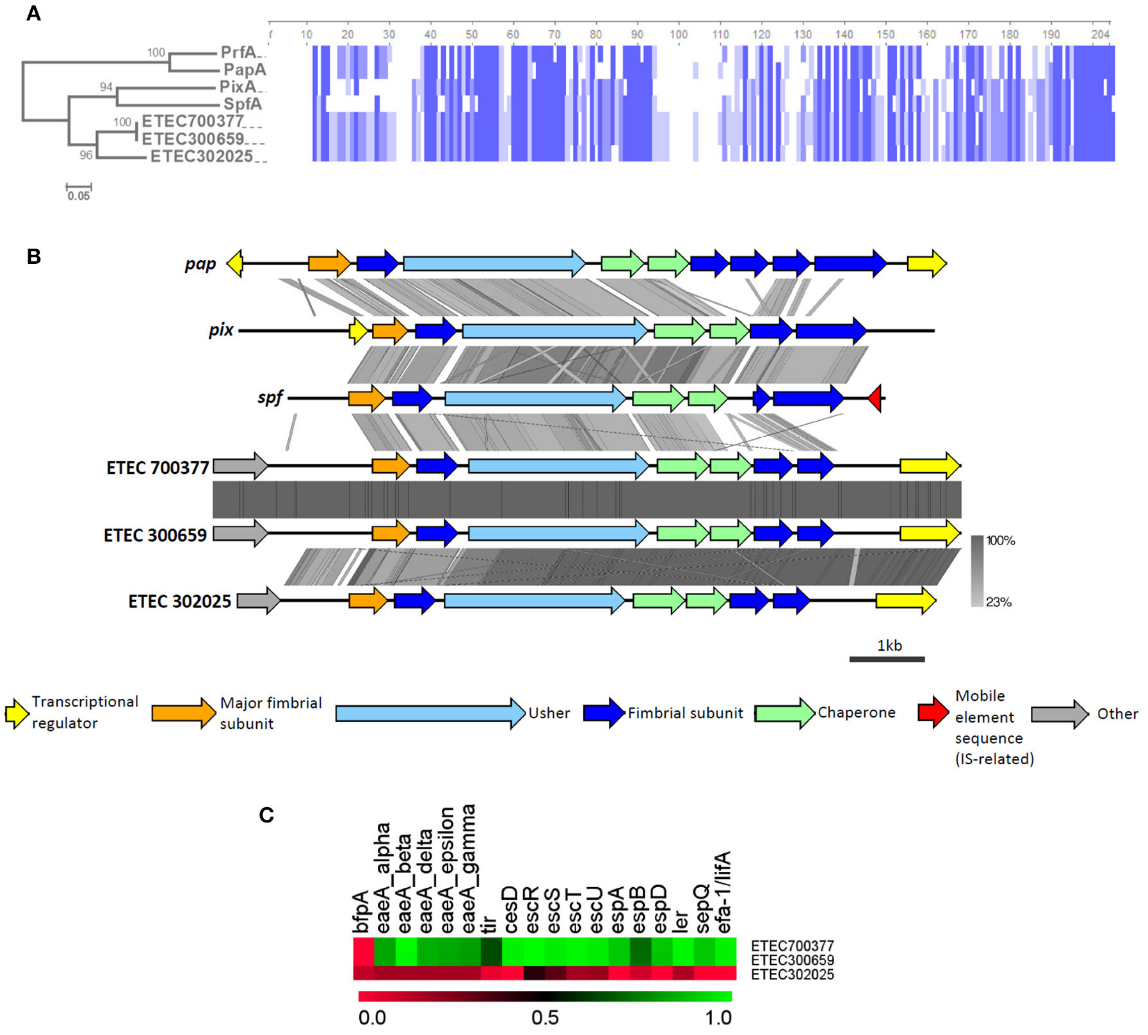


FIGURE 7 | Analysis of π-CU pili loci found in AN-ETEC strains. (A) Alignment of the amino acid sequences of putative π-CU pili major structural subunits of AN-ETEC strains with those of known π-CU pili (PrfA, SpfA, PapA, PixA, and SpfA). Blue blocks indicate identical regions and the color fades as the identity percentage decreases. The tree was inferred from alignment by the neighbor joining method with 1000 bootstrap replicates. **(B)** Alignment of the predicted π-CU pili loci found in AN-ETEC strains using tblastx in Easyfig. Known π-CU pili loci *prfA*, *papA*, *pixA*, and *spfA* were included. **(C)** Heat map derived from the screening of EPEC virulence genes in genomes of AN-ETEC strains harboring π-CU pili loci.

seems to bear two β-, γ_2-, κ-, or π-CU pili loci simultaneously, but they do harbor representatives of different families. In both, ETEC 100664 and ETEC 702251, β-, γ_2-, and κ-CU pili loci were identified; while seven strains carry β-, and γ_2-CU pili loci simultaneously (ETEC 8350a, ETEC 200144, ETEC 702332, ETEC 4155a, ETEC 9312a, ETEC 9343a, and ETEC 9788a). No β-, γ_2-, nor κ-CU pili loci were identified among the AN-ETEC strains carrying π-CU pili loci.

In the case of previously described loci, we maintained the established names; for novel gene variants, these names were complemented with subscripts numbers or letters. We now propose five novel loci names. Thus, the locus name *yhc* was maintained for identified β-CU pili loci, with novel variants of *yhcF* being named *yhcF*$_2$ and *yhcF*$_3$ (Korea et al., 2010). For the γ_2-CU pili loci, the previous names *crs/CS26*, *cma/CS27b*, *cnm/CS28a*, and *cnm/CS28b* were kept in the case of loci found in 14 AN-ETEC strains; and a novel variant of *crs/CS26* was identified (*crs/CS26b*) (Nada et al., 2011). Four novel names were given to the γ_2-CU pili loci found in nine strains: *ctg* (CU ETEC gamma-2 pilus locus), *gep* (gamma-2 ETEC pilus locus), *gte* (gamma-two pilus locus of ETEC) and *gtt* (gamma-two pilus locus of toxigenic *E. coli*). As κ-CU pili loci were found in three AN-ETEC strains and displayed high identity with the *aal* locus along their entire extension, the name *aal* was maintained and variants for *aalE* were introduced (*aalE*$_2$ and *aalE*$_3$; Del Canto et al., 2012). Finally, the acronym *ctp* (CU ETEC pi pilus locus) was chosen for the π-CU pili loci found in three AN-ETEC strains, with variants for the putative structural subunit genes *ctpA* and *ctpH* (*ctpA*$_2$ and *ctpH*$_2$).

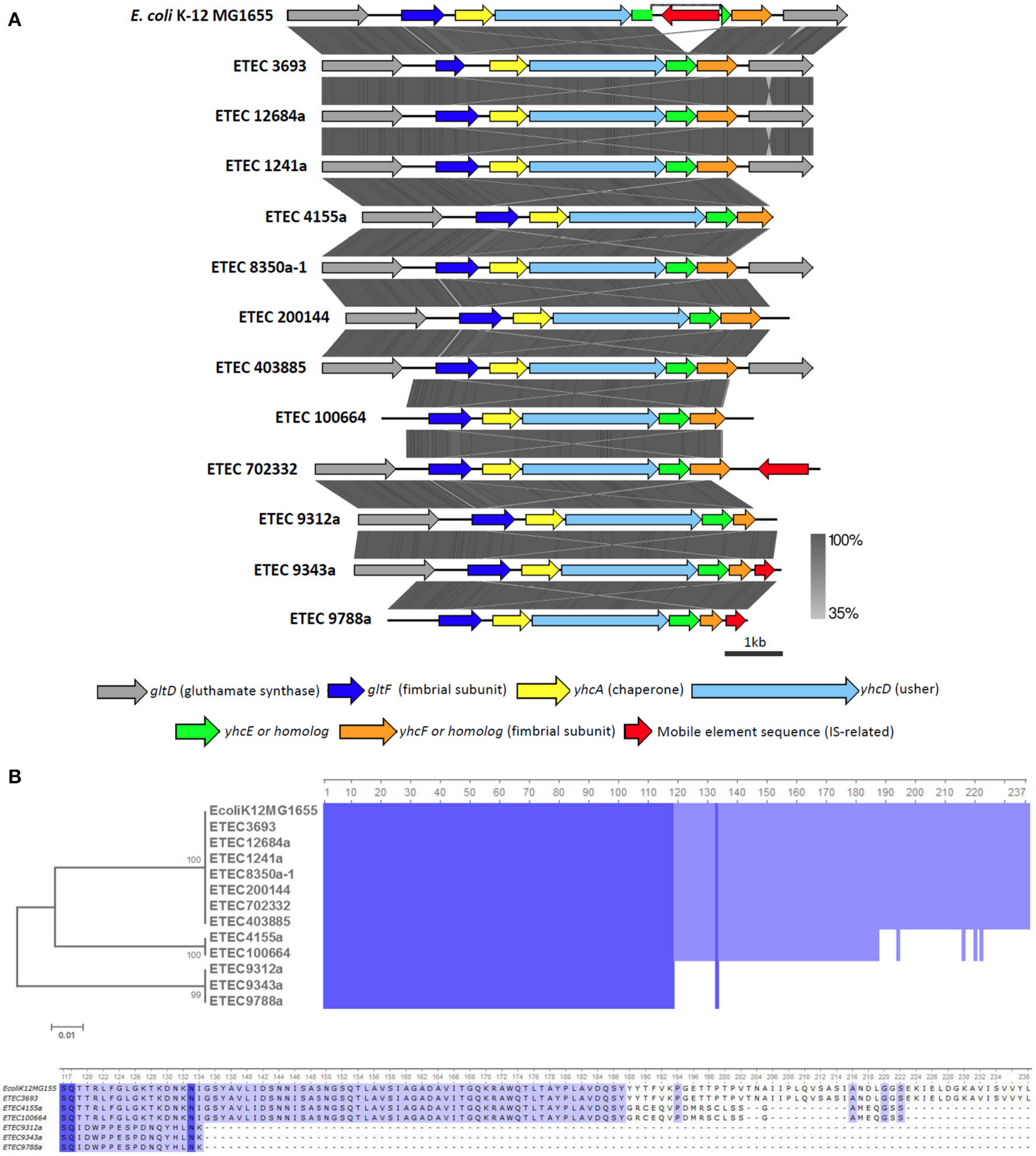
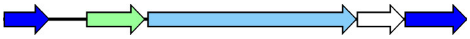
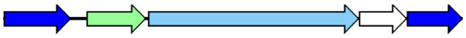
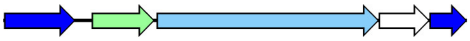
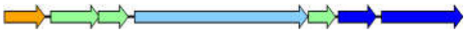
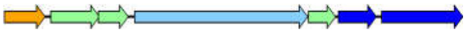
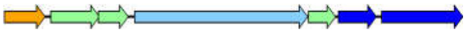
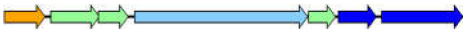
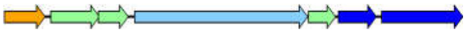
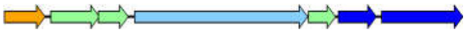
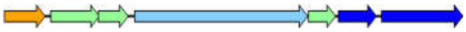
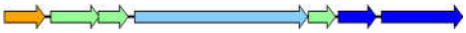
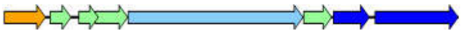
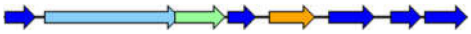
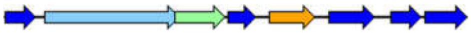
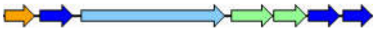
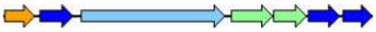

A
E. coli K-12 MG1655
ETEC 3693
ETEC 12684a
ETEC 1241a
ETEC 4155a
ETEC 8350a-1
ETEC 200144
ETEC 403885
ETEC 100664
ETEC 702332
ETEC 9312a
ETEC 9343a
ETEC 9788a
100%
35%
1kb
gltD (gluthamate synthase)
gltF (fimbrial subunit)
yhcA (chaperone)
yhcD (usher)
yhcE or homolog
yhcF or homolog (fimbrial subunit)
Mobile element sequence (IS-related)
B
1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 237
EcoliK12MG1655
ETEC3693
ETEC12684a
ETEC1241a
ETEC8350a-1
ETEC200144
ETEC702332
ETEC403885
ETEC4155a
ETEC100664
ETEC9312a
ETEC9343a
ETEC9788a
100
100
99
0.01
117 120 122 124 126 128 130 132 134 136 138 140 142 144 146 148 150 152 154 156 158 160 162 164 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194 196 198 200 202 204 206 208 210 212 214 216 218 220 222 224 226 228 230 232 234 238
EcoliK12MG155 SQTTRLFGLGKTKDNKNIGSYAVLIDSNNISASNGSQTLAVSIAGADAVITGQKRAWQTLTAYPLAVDQSYYYTFVKPGETTPTPVTNAIIPLQVSASIANDLGGSEKIELDGKAVISVVYL
ETEC3693 SQTTRLFGLGKTKDNKNIGSYAVLIDSNNISASNGSQTLAVSIAGADAVITGQKRAWQTLTAYPLAVDQSYYYTFVKPGETTPTPVTNAIIPLQVSASIANDLGGSEKIELDGKAVISVVYL
ETEC4155a SQTTRLFGLGKTKDNKNIGSYAVLIDSNNISASNGSQTLAVSIAGADAVITGQKRAWQTLTAYPLAVDQSYGRCEQVPDMRSCLSS--G----------AMEQGSS----------------
ETEC100664 SQTTRLFGLGKTKDNKNIGSYAVLIDSNNISASNGSQTLAVSIAGADAVITGQKRAWQTLTAYPLAVDQSYGRCEQVPDMRSCLSS--G----------AMEQGSS----------------
ETEC9312a SQIDWPPESPDNQYHLNK--
ETEC9343a SQIDWPPESPDNQYHLNK--
ETEC9788a SQIDWPPESPDNQYHLNK--

FIGURE 8 | Analysis of β-CU pili loci found in AN-ETEC strains. (A) Alignment of the predicted β-CU pili loci found in AN-ETEC strains using tblastx in Easyfig. The known β-CU locus of *E. coli* K-12 MG1655 was included. **(B)** Alignment of the YhcE amino acid sequences found in AN-ETEC strains. YhcE of *E. coli* K-12 MG1655 was included. Blue blocks indicate identical regions and the color fades as the identity percentage decreases. The tree was inferred from alignment using the neighbor joining method with 1000 bootstrap replicates. The lower panel shows a zoom in the variable C-terminal region of YhcE.

DISCUSSION

Understanding the infectious mechanisms of ETEC have been hampered by the fact that a significant portion of the ETEC isolates obtained from diarrhea cases worldwide have tested negative for currently known adhesins. Searching for and identifying novel adherence determinants can therefore contribute to the development of improved diagnostic methods and aid in the identification of potential targets for vaccines development and/or anti-adherence therapies. Massive sequencing and comparative genomics are valuable tools for these purposes (Sjöling et al., 2015), allowing the characterization of novel ETEC adhesin loci. We focused primarily on CU pili, as these are the main structures of currently known ETEC adhesins, particularly among CFs (Madhavan and Sakellaris, 2015). We chose a set of strains from different

TABLE 2 | Summary of the pili loci, or genes, identified in this work.

Locus (CU Family)		Locus organization	AN-ETEC strains	References
yhc (β)	*gltFyhcADF*		3693, 12684a, 1241a, 8350a-1, 200144, 702332, 403885	Korea et al., 2010
yhc (β)	*gltFyhcAD**F$_2$***		4155a, 100664	Korea et al., 2010; This work
yhc (β)	*gltFyhcAD**F$_3$***		9312a, 9343a, 9788a	Korea et al., 2010; This work
crs/CS26b (γ$_2$)	*crs**H$_b$**BCDEFG*		100664	Nada et al., 2011; This work
cma/CS27b (γ$_2$)	*cma**H$_b$**BCDEFG*		4155a, 603388, 1030c-1, 963a-1, 703200, 3693, 503210	Nada et al., 2011; This work
cnm/CS28a (γ$_2$)	*cnm**H$_a$**BCDEFG*		8350a-1, 200144, 702251	Nada et al., 2011; This work
cnm/CS28b (γ$_2$)	*cnm**H$_b$**BCDEFG*		200617, 505077, 8255a	Nada et al., 2011; This work
ctg (γ$_2$)	*ctgABCDEFG*		9312a, 9343a, 9788a, 401900	This work
ctg (γ$_2$)	*ctg**A$_2$**BCDEFG*		10380a	This work
gep (γ$_2$)	*gepABCDEFG*		15617a, 203518	This work
gte (γ$_2$)	*gteABCDEFG*		1241a	This work
gtt (γ$_2$)	*gttABCDEFG*		702332	This work
aal (κ)	*aalABCD**E$_2$**FGH*		402594	Del Canto et al., 2012; This work
aal (κ)	*aalABCD**E$_3$**FGH*		100664, 702251	Del Canto et al., 2012; This work
ctp (π)	*ctpAHCDEJF*		700377, 300659	Dutta et al., 2015; This work
ctp (π)	*ctp**A$_2$H$_2$**CDEJF*		302025	This work

🟧 *Major structural subunit* 🟦 *Structural subunit* ▭ *Usher* 🟩 *Chaperone* ⬜ *Unknown.*

Genetic variants are in bold letters.

geographical locations and belonging to different serogroups. Enterotoxin profiles were not a requisite for strain selection, but most of the AN-ETEC strains were indeed positive for LT and not for ST. This is consistent with features of previously reported CF negative ETEC strains (Wolf, 1997; Shaheen et al., 2009; Rivera et al., 2010; Del Canto et al., 2011). Also consistent with previous evidence, most of our AN-ETEC strain collection belonged to the A and B$_1$ phylogroups of *E. coli* (von Mentzer et al., 2014). This has been generally observed for ETEC strains, regardless of their virulence repertoire (von Mentzer et al., 2014). On the other hand, consistent with the diverse nature of ETEC strains, there were no predominant sequence types (von Mentzer et al., 2014).

As part of our goal, we aimed to find loci that were absent in non-pathogenic *E. coli*. Following this strategy, we identified 10 pili loci, distributed among 33 of the 35 AN-ETEC strains analyzed. Searching for genes encoding other adhesin types such as type IV pili, or proteins involved in their assembly, could allow the identification of additional loci, especially for cases such as the two strains resulting negative in our screening. In addition, these two strains may harbor CU pili homologous to those produced by non-pathogenic *E. coli*, a set that was excluded in our work. In this scenario, a detailed "case by case" analysis may be required to identify novel CU pili from alignment-based strategies, particularly when protein sequence identity fall to the 20–30% range, also known as the twilight zone (Blake and Cohen, 2001).

Representatives of the CU-pili families β, γ_2, κ, and π, were found and described. The most common CU-CFs worldwide belong to the α (CFA/I, CS1, CS2, CS4, CS5, CS14, CS17, and CS19) and γ_3 families (CS3 and CS6); these are the same CU-CFs that have been included in vaccine candidate formulations (O'Ryan et al., 2015; Zhang and Sack, 2015). CFs of the CU-families γ_2 and κ are not the most common in ETEC strains (Isidean et al., 2011). However, the identification of several γ_2-CU pili loci, described here and in previous works, suggests that they may be a numerous and important group among the ETEC adhesins. Certainly, wide screenings in ETEC collections are required to test this hypothesis. A search in databases using blastp, by introducing γ_2, κ, and π-PMS described here, identified a limited number of identical registries. $CrsH_b$ (γ_2, one hit), CtgA (γ_2, three hits), GepA (γ_2, three hits), GteA (γ_2, eight hits), and CtpA (π, three hits) were the only matched amino acid sequences with 100% identity and 100% coverage. In contrast, YhcE (β family) matched 353 registries, suggesting the absence of the insertion sequence that disrupts the *yhc* locus in *E. coli* K-12 strains. Upcoming massive sequencing projects and/or specific PCR screening will help to determine the global distribution of the loci identified here.

As sequencing projects are generating data to be deposited in databanks, assigning names to novel loci or proteins is not a trivial matter. We only assigned a name to a locus when, according to the number of genes it contains, it could direct the assembly of a potentially functional CU-pilus. We decided to keep the previously given names in the cases of the β-, κ-, as well as some of the γ_2-CU pili loci identified in this work, as these showed high percentages of identity with previously described loci (>95%). When some of the genes identified showed more dissimilarity to their known homologs, but the rest of the locus was identical, we still used the original names from the literature, using numbers to differentiate the novel variants. Novel acronyms were introduced for five of the γ_2- and the two of the π-CU pili loci. We have decided not to add novel representatives to the CF list until evidence of the existence of the pili has been confirmed and evaluation of their functional activity has been performed.

None of the CFs described to date belongs to the π family. Dutta et al. reported the presence of a homolog locus of *pap*, which encodes the P fimbria, in an ETEC strain that also has genes that used to be part of the LEE locus in atypical EPEC strains (Dutta et al., 2015). In the current project, we identified two π-CU pili loci, one of them identical to that described by Dutta et al. Our analysis was consistent in terms of the relatedness with the *pap* locus, but we found a higher identity with the *sfp* locus, which was first described in an enterohemorrhagic *E. coli* O157:H- strain, as is determinant in the capacity of agglutinating erythrocytes. These findings suggest that π-CU pili loci could also be part of the wide spectrum of ETEC adhesins. We estimated that percentages of identity to both the *pap* and *sfp* loci were not high enough to use one of these names for the novel ETEC π-CU pili loci. Therefore, we suggested the acronym *ctp*.

Among all the CU-pili families described in *E. coli* isolates, the β family is the one that has received the least attention. There are no published data on the production of β-CU pili or adhesin activity (Nuccio and Bäumler, 2007). In *E. coli* K-12 it is considered to be cryptic, because of the presence of an insertion sequence that would logically disrupt its functionality (Korea et al., 2010). However, the identification of non-disrupted β-CU pili loci in several AN-ETEC genomes, suggests that they may direct the assembly of functional structures. Two putative fimbrial subunits (GltF and YhcF), one chaperone (YhcA), and the usher (YhcD) are potentially encoded by β-CU pili loci, in addition to a protein with no predictable function (YhcE). A comprehensive review has suggested that, due to the lack of putative tip subunits, β-CU pili would be afimbrial structures or thin fibers (Nuccio and Bäumler, 2007).

Further research will be needed to prove the functional role of the CU pili loci identified in this work in ETEC pathogenesis. Preliminarily, we have obtained deletion mutants for a few representative ETEC strains and evaluated their capacity to attach to confluent monolayers of Caco-2 cells (**Figure S2** in supporting information). ETEC strains 100664 and 702332 lacking γ_2-CU loci *crs* and *gtt*, respectively, showed a reduced adhesion capacity compared to the wild type strains, even when this difference was only statistically significant in the case of *crs* (**Figure S2**). The same effect was noticed in ETEC strains 402594 and 302025 lacking the κ-CU locus *aal* (harboring $aalE_2$ variant) and π-CU locus *ctp* ($ctpA_2H_2$ variant). No differences were observed in mutants lacking β-CU locus *yhc*, γ_2-CU locus *cnm*, and π-CU locus *ctp*, compared to the wild types; while a significant increase was noticed in ETEC 9343a lacking the γ_2-CU locus *ctg* (**Figure S2**). These results suggest that some of the identified CU pili loci encode determinants of *in vitro* bacterial adherence. Testing of the proper expression conditions and also other cell lines or adhesion substrates, may be determinant for future functional evaluations of the CU pili identified here. In addition, as other kind of structures different from CU pili may be directing AN-ETEC adherence to Caco-2 cells, double mutants or recombinant expression in non-pathogenic *E. coli* strains could be required.

In conclusion, ETEC strains lacking currently known CFs bear genes that encode related structures, particularly CU-pili of the γ_2 and κ families. Beta- and π-CU pili are new members within the diverse ETEC adhesin repertoire. These findings add novel and valuable data to the large list of potential ETEC "weapons." Future investigation will characterize their specific role in pathogenicity and whether they are suitable antigens for vaccine candidate formulations.

AUTHOR CONTRIBUTIONS

FD: study design, strain characterization, data analysis, bioinformatic analysis, manuscript writing. MO: study design, data analysis, manuscript writing. MP: strain characterization, data analysis. AT: strain characterization, data analysis. DG: stain characterization, data analysis. RV: bioinformatic analysis. LC: bioinformatic analysis. AM: strain characterization, data analysis. RM: strain characterization, data analysis. DH: strain characterization, data analysis. BC: strain characterization, data analysis. MML: study design, data analysis. CMH: data analysis, bioinformatic analysis. MP: data analysis, bioinformatic analysis. DAR: data analysis, bioinformatic analysis, manuscript writing. OCS: study design, data analysis, bioinformatic analysis, manuscript writing. RV: study design, data analysis, manuscript writing.

FUNDING

This work was funded with the support of the following grants: Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT) grants 3130555 (FD), 11150966 (FD), 1110260 (RV), 1161161 (RV). Grant ID 38874, "Diarrheal disease in infants and young children in developing countries" and Grant ID 1016839 "Metagenomics Based Discovery of New Viral and Eukaryotic Pathogens Causing Diarrheal Disease" from the Bill and Melinda Gates Foundation.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ann-Mari Svennerholm and Gudrun Wiklund for providing us with control strains for the detection of CS8, CS13, CS15, CS18, and CS22. Scientific writing support was provided by Dr. Lili Fox Velez.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcimb.2016.00200/full#supplementary-material

Figure S1 | (A) Alignment of nucleotide sequences of the gene encoding a putative type-IV pilin, found in ETEC 401909, with genes cofA and lngA encoding major subunits of CS8 and CS21, respectively. The internal regions shown correspond to those recognized by primers (forward, left; reverse, right) used in CFs detection. **(B)** Alignment of amino acid sequences of the pilins of CS8, CS21, CS8b, and that found in ETEC 401909.

Figure S2 | Evaluation of the adherence capacity of ETEC strains to Caco-2 cells, after knocking out CU pili loci identified in this work. (A) β-CU: Locus *yhc* in ETEC 100664, **(B)** γ_2-CU: locus *crs* in ETEC 100664, locus *cnm* in ETEC 8350a-1, locus *crs* in ETEC 9343a, and locus *gtt* in ETEC 702332 (all loci belonging to family), **(C)** κ-CU: locus *aal* in ETEC 402594, **(D)** π-CU: locus *ctp* in ETEC 300659, and locus *ctp* in ETEC 302025. The graphics show the number (average ± *SEM*) of cell-associated colony forming units (CFU) recovered after 3 h of infection. Mutant strains were obtained by allelic exchange mediated by the lambda red recombinase system using pKD46 or pSIMs plasmids (Datsenko and Wanner, 2000; Sharan et al., 2009) and the primers listed in the table below. Adherence capacity was evaluated by infecting confluent Caco-2 monolayers with a multiplicity of 10 bacteria per cell, for 3 h at 37°C in a 5% CO_2 atmosphere. After repeated washes with saline buffer, cells were lysed with 0.1% Triton X-100. The suspension was serially diluted and seeded onto LB agar plates for CFU counting. Data were analyzed using unpaired *t*-test with Welch's correction and differences were considered significant when $P < 0.05$.

Table S1 | Accession codes for gene sequences used in this work.

Table S2 | PCR primers used for CF gene detection.

Table S3 | Chromosome and plasmid sequences of *Escherichia coli* strains accessed in this work.

Data Sheet 1 | Putative major subunits encoded by CU-pili loci identified in this work.

REFERENCES

Ahmed, T., Bhuiyan, T. R., Zaman, K., Sinclair, D., and Qadri, F. (2013). Vaccines for preventing enterotoxigenic *Escherichia coli* (ETEC) diarrhoea. *Cochrane Database Syst. Rev.* 7:CD009029. doi: 10.1002/14651858.CD009029

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12:402. doi: 10.1186/1471-2164-12-402

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., and Edwards, R. A. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Blake, J. D., and Cohen, F. E. (2001). Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.* 307, 721–735. doi: 10.1006/jmbi.2001.4495

Brunder, W., Khan, A. S., Hacker, J., and Karch, H. (2001). Novel type of fimbriae encoded by the large plasmid of sorbitol-fermenting enterohemorrhagic *Escherichia coli* O157:H(-). *Infect. Immun.* 69, 4447–4457. doi: 10.1128/IAI.69.7.4447-4457.2001

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30, 31–37. doi: 10.1093/bioinformatics/btt310

Clermont, O., Christenson, J. K., Denamur, E., and Gordon, D. M. (2013). The Clermont *Escherichia coli* phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* 5, 58–65. doi: 10.1111/1758-2229.12019

Croxen, M. A., Law, R. J., Scholz, R., Keeney, K. M., Wlodarska, M., and Finlay, B. B. (2013). Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin. Microbiol. Rev.* 26, 822–880. doi: 10.1128/CMR.00022-13

Del Canto, F., Botkin, D. J., Valenzuela, P., Popov, V., Ruiz-Perez, F., Nataro, J. P., et al. (2012). Identification of coli surface antigen 23, a novel adhesin of enterotoxigenic *Escherichia coli*. *Infect. Immun.* 80, 2791–2801. doi: 10.1128/IAI.00263-12

Del Canto, F., Valenzuela, P., Cantero, L., Bronstein, J., Blanco, J. E., Blanco, J., et al. (2011). Distribution of classical and nonclassical virulence genes in enterotoxigenic *Escherichia coli* isolates from Chilean children and tRNA gene screening for putative insertion sites for genomic islands. *J. Clin. Microbiol.* 49, 3198–3203. doi: 10.1128/JCM.02473-10

Dutta, S., Pazhani, G. P., Nataro, J. P., and Ramamurthy, T. (2015). Heterogenic virulence in a diarrheagenic *Escherichia coli*: evidence for an EPEC expressing heat-labile toxin of ETEC. *Int. J. Med. Microbiol.* 305, 47–54. doi: 10.1016/j.ijmm.2014.10.006

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Isidean, S. D., Riddle, M. S., Savarino, S. J., and Porter, C. K. (2011). A systematic review of ETEC epidemiology focusing on colonization factor and toxin expression. *Vaccine* 29, 6167–6178. doi: 10.1016/j.vaccine.2011.06.084

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M., and Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS ONE* 9:e104984. doi: 10.1371/journal.pone.0104984

Korea, C. G., Badouraly, R., Prevost, M. C., Ghigo, J. M., and Beloin, C. (2010). *Escherichia coli* K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. *Environ. Microbiol.* 12, 1957–1977. doi: 10.1111/j.1462-2920.2010.02202.x

Kotloff, K. L., Nataro, J. P., Blackwelder, W. C., Nasrin, D., Farag, T. H., Panchalingam, S., et al. (2013). Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet* 382, 209–222. doi: 10.1016/S0140-6736(13)60844-2

Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., and Marvig, R. L. (2012). Multilocus sequence typing of total genome sequenced bacteria. *J. Clin. Micobiol.* 50, 1355–1361. doi: 10.1128/JCM.06094-11

Lukjancenko, O., Wassenaar, T. M., and Ussery, D. W. (2010). Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* 60, 708–720. doi: 10.1007/s00248-010-9717-3

Madhavan, T. P., and Sakellaris, H. (2015). Colonization factors of enterotoxigenic *Escherichia coli*. *Adv. Appl. Microbiol.* 90, 155–197. doi: 10.1016/bs.aambs.2014.09.003

Nada, R. A., Shaheen, H. I., Khalil, S. B., Mansour, A., El-Sayed, N., Touni, I., et al. (2011). Discovery and phylogenetic analysis of novel members of class b enterotoxigenic *Escherichia coli* adhesive fimbriae. *J. Clin. Microbiol.* 49, 1403–1410. doi: 10.1128/JCM.02006-10

Njoroge, S. M., Boinett, C. J., Madé, L. F., Ouko, T. T., Fèvre, E. M., Thomson, N. R., et al. (2015). A putative, novel coli surface antigen 8B (CS8B) of enterotoxigenic *Escherichia coli*. *Pathog. Dis.* 73:ftv047. doi: 10.1093/femspd/ftv047

Nuccio, S. P., and Bäumler, A. J. (2007). Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol. Mol. Biol. Rev.* 71, 551–575. doi: 10.1128/MMBR.00014-07

Okonechnikov, K., Golosova, O., Fursov, M., and UGENE team. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

O'Ryan, M., Vidal, R., Del Canto, F., Salazar, J., and Montero, D. (2015). Vaccines for viral and bacterial pathogens causing acute gastroenteritis: Part II: vaccines for Shigella, Salmonella, enterotoxigenic *E. coli* (ETEC) enterohemorragic *E. coli* (EHEC) and *Campylobacter jejuni*. *Hum. Vaccin. Immunother.* 11, 601–619. doi: 10.1080/21645515.2015.1011578

Panchalingam, S., Antonio, M., Hossain, A., Mandomando, I., Ochieng, B., Oundo, J., et al. (2012). Diagnostic microbiologic methods in the GEMS-1 case/control study. *Clin. Infect. Dis.* 55(Suppl. 4), S294–S302. doi: 10.1093/cid/cis754

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114

Pichel, M., Binsztein, N., and Viboud, G. (2000). CS22, a novel human enterotoxigenic *Escherichia coli* adhesin, is related to CS15. *Infect. Immun.* 68, 3280–3285. doi: 10.1128/IAI.68.6.3280-3285.2000

Qadri, F., Svennerholm, A. M., Faruque, A. S., and Sack, R. B. (2005). Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. *Clin. Microbiol. Rev.* 18, 465–483. doi: 10.1128/CMR.18.3.465-483.2005

Rivera, F. P., Ochoa, T. J., Maves, R. C., Bernal, M., Medina, A. M., Meza, R., et al. (2010). Genotypic and phenotypic characterization of enterotoxigenic *Escherichia coli* strains isolated from Peruvian children. *J. Clin. Microbiol.* 48, 3198–3203. doi: 10.1128/JCM.00644-10

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., et al. (2006). TM4 microarray software suite. *Methods Enzymol.* 411, 134–193. doi: 10.1016/S0076-6879(06)11009-5

Sahl, J. W., Caporaso, J. G., Rasko, D. A., and Keim, P. (2014). The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. *Peer J.* 2:e332. doi: 10.7717/peerj.332

Shaheen, H. I., Abdel Messih, I. A., Klena, J. D., Mansour, A., El-Wakkeel, Z., Wierzba, T. F., et al. (2009). Phenotypic and genotypic analysis of enterotoxigenic *Escherichia coli* in samples obtained from Egyptian children presenting to referral hospitals. *J. Clin. Microbiol.* 47, 189–197. doi: 10.1128/JCM.01282-08

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Sjöling, Å., von Mentzer, A., and Svennerholm, A. M. (2015). Implications of enterotoxigenic *Escherichia coli* genomics for vaccine development. *Expert. Rev. Vaccines* 14, 551–560. doi: 10.1586/14760584.2015.996553

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Turner, S. M., Scott-Tucker, A., Cooper, L. M., and Henderson, I. R. (2006). Weapons of mass destruction: virulence factors of the global killer enterotoxigenic *Escherichia coli*. *FEMS Microbiol. Lett.* 263, 10–20. doi: 10.1111/j.1574-6968.2006.00401.x

Valvatne, H., Sommerfelt, H., Gaastra, W., Bhan, M. K., and Grewal, H. M. (1996). Identification and characterization of CS20, a new putative colonization factor of enterotoxigenic *Escherichia coli*. *Infect. Immun.* 64, 2635–2642.

von Mentzer, A., Connor, T. R., Wieler, L. H., Semmler, T., Iguchi, A., Thomson, N. R., Rasko, D. A., et al. (2014). Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat. Genet.* 46, 1321–1326. doi: 10.1038/ng.3145

Waksman, G., and Hultgren, S. J. (2009). Structural biology of the chaperone-usher pathway of pilus biogenesis. *Nat. Rev. Microbiol.* 7, 765–774. doi: 10.1038/nrmicro2220

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., and Wieler, L. H. (2006). Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.* 60, 1136–1151. doi: 10.1111/j.1365-2958.2006.05172.x

Wolf, M. K. (1997). Occurrence, distribution, and associations of O and H serogroups, colonization factor antigens, and toxins of enterotoxigenic *Escherichia coli*. *Clin. Microbiol. Rev.* 10, 569–584.

Wurpel, D. J., Beatson, S. A., Totsika, M., Petty, N. K., and Schembri, M. A. (2013). Chaperone-usher fimbriae of *Escherichia coli*. *PLoS ONE* 8:e52835. doi: 10.1371/journal.pone.0052835

Zhang, W., and Sack, D. A. (2015). Current progress in developing subunit vaccines against enterotoxigenic *Escherichia coli*-associated diarrhea. *Clin. Vaccine Immunol.* 22, 983–991. doi: 10.1128/CVI.00224-15

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.