Whole genome sequencing of enterotoxigenic *Escherichia coli* (ETEC)

Identification of ETEC lineages and novel colonization factors

Astrid von Mentzer

Department of Microbiology and Immunology
Institute of Biomedicine
Sahlgrenska Academy at University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2016

Cover illustration: Astrid von Mentzer

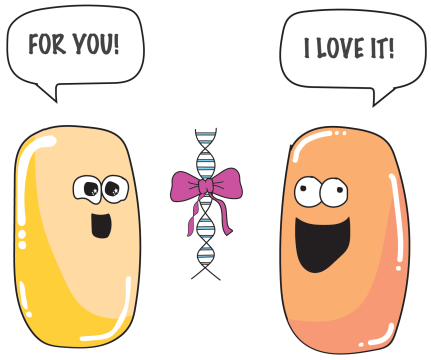
Whole genome sequencing of enterotoxigenic *Escherichia coli* (ETEC)

astrid.von.mentzer@gu.se

ISBN: 978-91-628-9800-7 (PRINT)
ISBN: 978-91-628-9801-4 (PDF)
http://hdl.handle.net/2077/41845

Printed in Gothenburg, Sweden 2016
Ineko AB

Till min familj ♥



Whole genome sequencing of enterotoxigenic *Escherichia coli* (ETEC)

Identification of ETEC lineages and novel colonization factors

Astrid von Mentzer
Department of Microbiology and Immunology, Institute of Biomedicine
Sahlgrenska Academy at University of Gothenburg
Gothenburg, Sweden

ABSTRACT

Enterotoxigenic *Escherichia coli* (ETEC) infections are a common cause of diarrhea but little is known about the evolution and genomic composition of ETEC. The main aim of this thesis was to generate a large collection of whole genome sequenced ETEC isolates to study the evolution of such bacteria on a global level and to search for novel colonization factors (CFs) using both classical and cutting-edge genomic techniques.

Using whole genome sequencing and epidemiological data of 362 human ETEC isolates collected during three decades from all over the world, we studied the population structure of ETEC. We could show that major ETEC lineages comprising isolates with specific virulence profiles, i.e. CFs and LT and/or ST toxins, are stable and spread worldwide. These findings suggest that the virulence genes have been acquired once and then spread through clonal expansion and that a vaccine based on the most prevalent CFs could be protective against a large proportion of ETEC diarrhea cases.

At least 30% of all clinical ETEC isolates lack a known CF. Therefore, we examined whole genome sequences of 94 "CF negative" isolates with the aim to identify novel CFs using two different approaches. **I)** By comparative genomics we have characterized a novel CF, CS30, which is related to the porcine CF 987P (F6). The major subunit of CS30 is 18.5 kD in size and the assembly of the fimbriae is dependent on its expression. CS30-positive bacteria are heavily fimbriated, as shown by electron microscopy, which promotes binding to human intestinal (Caco-2) cells. Furthermore, CS30 expression is thermo-regulated. **II)** By means of phenotypic analyses (SDS-PAGE, Caco-2 adhesion assays and electron microscopy) we have identified a number of isolates that may harbor additional putative fimbrial and non-fimbrial novel CFs. Nine candidate isolates with putative novel CFs were identified from 35 isolates: these were shown to express thermo-regulated proteins of 12-25 kD by SDS-PAGE analyses indicating the presence of major subunits and these isolates were found to bind well to Caco-2 cells. Based on further analyses of these isolates, using comparative genomics to identify CF related genes/operons, four AFA/Dr/AAF-like operons and an operon related to the porcine CF K88, were identified. The findings in this thesis have improved the knowledge of ETEC genomics and will provide a basis for future studies of ETEC transmission and pathogenicity.

Keywords: ETEC, whole genome sequencing, evolution, colonization factor, reverse genetics

ISBN: 978-91-628-9800-7 (PRINT) **ISBN:** 978-91-628-9801-4 (PDF)

SAMMANFATTNING PÅ SVENSKA

Enterotoxinbildande *Escherichia coli* bakterier (ETEC) är en av de främsta orsakerna till diarré hos barn och vuxna i utvecklingsländer. Det är även den vanligaste orsaken till turistdiarré. ETEC bakterier uttrycker ytproteiner, kolonisationsfaktorer, som möjliggör bindning till tarmepitelet samt utsöndrar toxiner (LT och ST: STh eller STp). Trots att ETEC är ett stort problem världen över så vet man inte tillräckligt om dess genetiska komposition.

Vi har skapat den största kollektionen av hel-genom sekvenserade ETEC, bestående av 362 stammar som har isolerats världen över mellan 1980 till 2011. Dessa stammar härstammar från barn och vuxna i endemiska områden samt resenärer och soldater med diarré. Genom att studera evolutionen av ETEC har vi kunnat visa att ETEC stammar grupperas i globalt spridda kloner. Stammar inom de största grupperna bär på samma kolonisationsfaktor- och toxin-gener, de delar alltså samma virulensprofil. Vi visar också att dessa grupper tycks vara stabila över tid och uppstod förmodligen i modern tid (ca 1837-1961) för att sedan spridas över världen.

Sedan tidigare vet man att minst 30 % av alla kliniskt isolerade ETEC stammar saknar en känd kolonisationsfaktor. Vi har använt s.k. reverse genetics (genomisk analys för att hitta kandidatgener) som ett sätt att hitta möjliga nya kolonisationsfaktorer hos stammar som saknat en känd kolonisationsfaktor. En ny kolonisationsfaktor, CS30, identifierades och karakteriserades. CS30 tillhör Class 1b gruppen av kolonisationsfaktorer hos humana ETEC och likt andra kolonisationsfaktorer i gruppen har CS30 en rigid fimbrie struktur som är temperaturreglerad. Vi har även använt oss av den tredje generationens sekvenseringsteknik, PacBio, och identifierat den plasmid som bär på CS30 operonet samt båda toxingenerna (LT och STp). CS30 visades vara relaterad till en 987P, en kolonisationsfaktor som finns hos gris-ETEC.

Vi har även genom fenotypiska metoder följt av genomiska analyser identifierat flera stammar med potentiella kolonisationsfaktorer. Proteinanalys, cellbindningsstudier och elektronmikroskopi användes för att välja ut vilka stammar som sedan skulle analyseras vidare med genomiska metoder. Den genotypiska analysen påvisade flera operon i de utvalda stammarna som potentiellt skulle kunna uttrycka kolonsationsfaktorer; dessa utgjordes av fyra olika operon relaterade till AFA/Dr/AAF familjen (hittas hos enteroaggregativa, uropatogena och ”diffusively adherent” *E. coli*) samt ett operon som är likt K88 (F4), en kolonisationsfaktor med gris-specificitet.

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals:

I **Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution**
von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm A-M, Sjöling Å and Dougan G
Nature Genetics, 2014. 46:1321-26.

II **Identification and characterization of a novel colonization factor based on whole genome sequencing in enterotoxigenic *Escherichia coli* (ETEC)**
von Mentzer A, Tobias J, Wiklund G, Aslett M, Dougan G, Sjöling Å and Svennerholm A-M
Submitted, 2016.

III **Identification of novel enterotoxigenic *Escherichia coli* (ETEC) putative colonization factors based on phenotypic and genotypic analyses**
von Mentzer A, Wiklund G, Dougan G, Sjöling Å and Svennerholm A-M
In manuscript.

CONTENT

ABBREVIATIONS

AAF	Aggregative adherence fimbriae
AC	Adenylate cyclase
AFA	Adherence fibrillar adhesin
AIEC	Adherent invasive *Escherichia coli*
AVG	Anti-virulence gene
BAPS	Bayesian analysis of the population structure
BLASTn	Basic local alignment search tool using a nucleotide query
BEAST	Bayesian evolutionary analysis sampling trees
cAMP	Cyclic adenosine monophosphate
cGMP	Cyclic guanosine monophosphate
CDS	Coding DNA sequence
CF	Colonization factor
CFTR	Cystic fibrosis transmembrane regulator
CFU	Colony forming unit
CS	Coli surface antigen
CU	Chaperone/Usher pathway
DAEC	Diffusively adherent *Escherichia coli*
EAEC	Enteroaggregative *Escherichia coli*
E. coli	*Escherichia coli*
EHEC	Enterohemmorraghic *Escherichia coli*
EIEC	Enteroinvasive *Escherichia coli*
EPEC	Enteropathogenic *Escherichia coli*
ER	Endoplasmic reticulum
ETEC	Enterotoxigenic *Escherichia coli*
ExPEC	Extraintestinal pathogenic *Escherichia coli*
GM1	monosialotetrahexosylganglioside
G_s	GTP-binding protein
HGT	Horizontal gene transfer
IS	Insertion sequence
L1-L21	Lineage 1 - Lineage 21
Le^a	Lewis antigen A
Lpf	Long polar fimbriae
LPS	Lipopolysaccharide

LT	Heat-labile toxin
MCG	Maximum common genome
MLST	Multi-locus sequence type
MRCA	Most recent common ancestor
NaCl	Sodium chloride
oriT	Origin of transfer
ORF	Open reading frame
(Q)MS	(Quantitative) Mass spectrometry
SMRT	Single-molecule real-time
SNP	Single nucleotide polymorphism
ST/STh/STp	Heat-stable toxin (STh: human; STp:porcine)
T2SS	Type II secretion system
TEM	Transmission electron microscopy
UPEC	Uropathogenic *Escherichia coli*
WGS	Whole genome sequencing

INTRODUCTION

Pathogenic *Escherichia coli (E. coli)* may arise by the acquisition of virulence features, mainly by mobile elements such as transposons, insertion sequences, bacteriophages and plasmids, which can either be integrated into the chromosome or be self-replicating within the host [1]. Pathogenic *E. coli* are divided into to groups based on the site of the infection: intestinal pathogenic *E. coli* causes diarrhea-associated infections and extra-intestinal pathogenic *E. coli* (ExPEC) causes disease outside of the intestine, such as urinary infections, meningitis and sepsis [2]. There are five major diarrhea-causing *E. coli*: enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC), enteroaggregative *E. coli* (EAEC), enterohemorraghic *E. coli* (EHEC) and enteroinvasive *E. coli* (EIEC) [3]. In addition, diffusively adherent *E. coli* (DAEC) may be considered a diarrhea-causing *E. coli*, although there are no uniform markers for its detection [1], and adherent invasive *E. coli* (AIEC) is a recently discovered pathotype that may be involved in Crohn's disease [1]. The site and mechanism of colonization, pathogenicity, clinical symptoms and outcomes can differ between these enteropathogenic *E. coli*, demonstrating their diversity [3].

Diarrhea is a leading cause of child morbidity and mortality in low and middle-income countries [4], where ETEC is a major cause of diarrhea, especially in young children, but also in adults and travelers visiting such endemic areas [5].

ETEC

Epidemiology

ETEC infection is common in low and middle-income countries and spreads via the fecal-oral route by contaminated food and/or water in settings of poor sanitation and inadequate drinking facilities. Both children and adults are at risk of getting infected, however the most vulnerable group are children below five years of age, who may suffer from multiple diarrheal episodes each year [5,6]. The symptoms from an ETEC infection are similar to those caused by many other enteric pathogens, and together with the lack of wide-spread simple diagnostic tools, makes estimating the global incidence of ETEC infections

difficult. However, in 2010, the Global Burden of Disease study estimated that ETEC yearly caused 120,000 deaths in both children and adults, with 38,000 deaths in children below four years of age [7]. ETEC is also the major cause of diarrhea in travelers visiting endemic areas, and is estimated to cause 10 million diarrhea episodes per year and accounts for approximately 15-35%, depending on geographical region, of all travelers' diarrhea [5,8,9]. Furthermore, ETEC is also a significant cause of infection in military personnel in the field [10].

ETEC also causes neonatal enteric infections in agricultural animals including, pigs, cattle and sheep. However, such strains seem to be animal specific and have not been found to infect humans [11].

Disease and treatment

ETEC infection ranges from mild to severe cholera-like diarrhea [5]. The diarrhea is usually self-limited, lasting three to four days. The infection may be treated by water and electrolyte rehydration to balance the loss of fluids and ions. The development of an ETEC vaccine is of key importance to prevent children and adults from getting infected. A multi-valent oral ETEC vaccine (recently named ETVAX) has been developed at the University of Gothenburg, which is evaluated in several human trials [12].

Pathogenesis

ETEC is transmitted through the fecal-oral route and may be ingested via contaminated food or water. The relatively high infectious dose of 10^6-10^{10} colony forming units (CFU) is required to cause diarrhea [13]. Children, elderly and immunosuppressed individuals may not require such large amounts of bacteria. Upon digestion, the bacteria colonizes the small intestine via expression of colonization factors (CFs), which are outer-membrane structures that adhere to the intestinal epithelium and are host-specific [11,14]. When in close proximity to the epithelial cells, the bacteria secrete enterotoxin(s), heat-labile toxin (LT) and/or heat-stable toxin (ST), causing ions and water efflux resulting in rapid onset watery diarrhea [5,15] allowing further spread of the bacteria in the environment. The mechanisms by which LT and ST cause diarrhea are described in **Figure 1**.

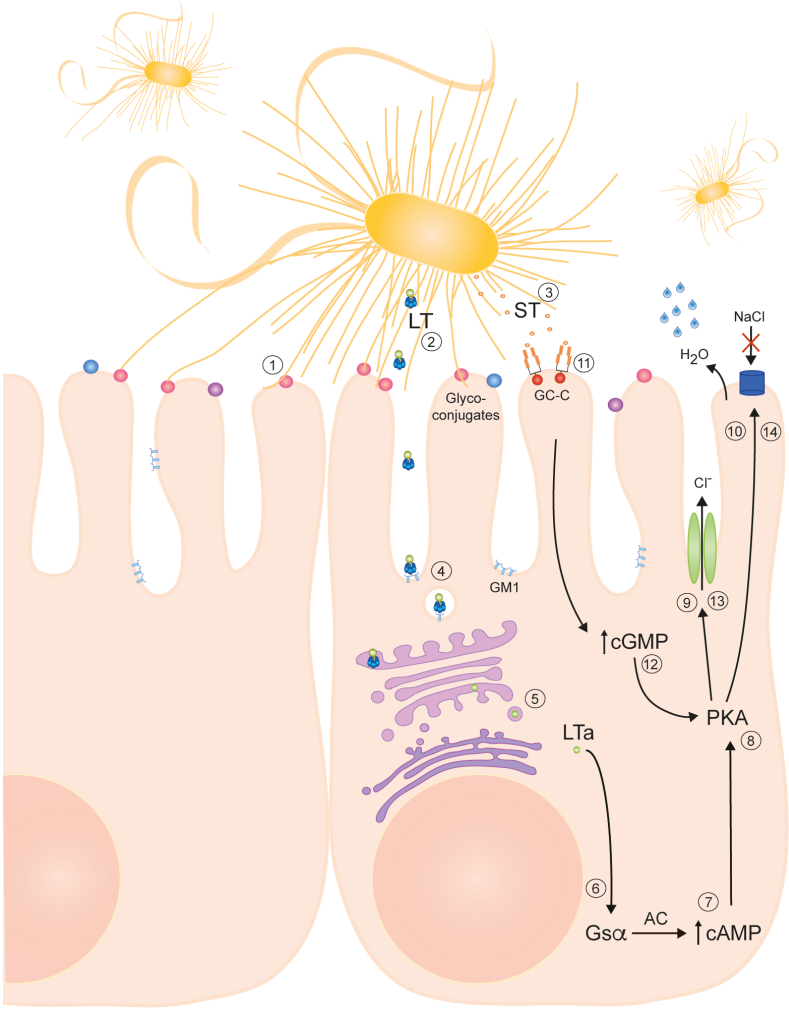


Fig 1. Pathogenesis of ETEC. 1. ETEC colonizes the small intestine via CFs adhering to glycoconjugates [14]. **2. & 3**. Enterotoxins (LT and/or ST) are secreted. **4.** LT binds to GM1 and other glycoproteins [16]. **5.** The holoxtoxin is endocytosed and travels to the endoplasmic reticulum (ER) via a retrograde pathway. The catalytic part of LTa is released and transported to the cytoplasm [16]. **6.** LTa ADP-ribosylates the alpha subunit of the GTP-binding protein (G_s), stimulating adenylate cyclase (AC). **7.** Leading to increased levels of cyclic AMP (cAMP). **8.** cAMP activates protein kinase (PKA), and **9.** phosphorylates the cystic fibrosis transmembrane regulator (CFTR) resulting in efflux of chloride ions and bicarbonate and **10.** blocks the uptake of sodium chloride (NaCl) [16]. **11.** ST binds to guanylate cyclase C, activating the inner membrane domain [17]. **12**. This results in elevated levels of cyclic GMP (cGMP) which activates protein kinase II, **13.** which activates CFTR in the same manner as PKA and **14.** blocks NaCl uptake. [18]. The effect of both LT and ST is an efflux of negatively charged ions and water, and a blockage of NaCl uptake.

Heat-labile and heat-stable enterotoxins

ETEC is defined by the production of one or two enterotoxins: LT and ST [5,16]. There are two types of LT: LT-I and LT-II, where LT-I is expressed by ETEC isolates infecting both humans and pigs and LT-II is found primarily in animal ETEC isolates. LT is structurally, immunologically and functionally closely related, but not identical to CT [19]. In this thesis, LT refers to LT-I. There are two variants of LT-I called LTh and LTp, known to infect humans and pigs, respectively [16]. In our studies, all isolates expressed LTh. LT is encoded by *eltA* and *eltB*, and are located on plasmids which may also harbor the gene *estA* encoding ST and/or genes encoding CFs. ST is also subdivided into the two major groups STa and STb, where STa is produced by isolates infecting humans and STb is only found in isolates infecting pigs and cattle [20,21]. There are also two types of STa: STh and STp, where the former is found in isolates infecting only humans and the later was originally identified in isolates from infected pigs. However, STp was later also identified in isolates collected from infected humans with diarrheal disease [22]. Both LT and ST are transported across the bacterial plasma membrane to the periplasm in a Sec-dependent manner [23], directed by N-terminal signalpeptides. LT is exported through the type II secretion system (T2SS), which is also called the general secretion pathway and is also present in other Gram-negative pathotypes [23], and ST is exported by the protein exporter TolC [24]. While most LT is trapped in the periplasm [25,26] ST is secreted to a greater extent. The mechanisms by which of LT and ST cause diarrhea are described above (see **Fig 1**).

Serotypes

O antigens are carbohydrates part of the lipopolysaccharide (LPS) expressed on the outer membrane. O antigens are composed of repeated subunits with great variation. At least 78 different O antigen serogroups have been identified in ETEC [27]; in addition, some ETEC isolates lack (rough strains) or express an as of yet unknown O antigen. Although O antigens are immunogenic their great variability makes an ETEC vaccine based solely on O antigens unfeasible.

Colonization factors

E. coli and other pathogenic Gram-negative bacteria often express outer membrane structures that may be used for adherence to the target host cell. One major assembly class is chaperone/usher (CU)-dependent pathways, including the classical and alternative pathways. ETEC CFs, Type I fimbriae and the related adherence fibrillar adhesin (AFA), Dr-adhesins and aggregative adherence fimbriae (AAF) (expressed by some EAEC, UPEC and DAEC) are all classified as CU structures [28,29]. They require chaperone(s) and an usher for assembly, which are part of an operon together with structural subunit(s), major and/or minor subunits. CU structures may be divided into specific groups based on the presence of conserved protein motifs in the usher and chaperone [28] or based on other features, such as antigenicity and operon structure [14]. ETEC CFs may be of fimbrial, fibrillar or non-fimbrial structures with the potential to mediate adherence to the human intestinal mucosa [14].

Assembly of CFs

Type I fimbriae and P pili are the most studied adhesins with regard to assembly. In general, assembly and structural proteins are transported across the inner membrane in a Sec-dependent manner [30-32]. The structural subunits (major and minor) interact with their respective chaperone for correct folding. The chaperone/subunit complex interacts with the outer membrane usher, which facilitates their secretion, with the most distal part being secreted first [33,34]. Some operons encode non-fimbrial structures that coil up on the cell surface, but most CF operons are assembled as fimbrial or fibrillar structures [14].

Known ETEC CFs

At least 25 phenotypically and epidemiologically different ETEC CFs (**Table 1**) have been identified so far [14,35,36]. These CFs can be divided into groups based on specific attributes, most commonly genetic and antigenic relatedness (**Fig 2** and **Table 1**). Three major groups have been classified: the CFA/I-like group, including CFA/I, CS1, CS2, CS4, CS14, CS17, CS19 and PCFO71 [37]: the CS5-like group, which including CS5 and the closely related CS7, and the Class 1b group: including the previously described CFs: CS12, CS18 and CS20, as well as the recently identified novel CFs CS26-CS28 [38] and CS30 (Paper II). The majority of Class 1b CFs are expressed by STp-positive ETEC

Table 1: List of ETEC CFs and their characteristics.

CF	Morphology[1]		Size (kD) [14][2]	Accession number	Ref
CFA/-like group					
CFA/I	F	7 nm	25.0	M55661.1	[39,40]
CS1	F	7 nm	15.2	AY536429.1	[41-43]
CS2	F	7 nm	15.4	Z47800	[41,43,44]
CS4	F	6 nm	15.0	AF296132.1	[45]
CS14	F	7 nm	15.0/15.5	AY283611	[46]
CS17	F	7 nm	15.5	AY515609.1	[47]
CS19	F	7 nm	15.0	AY288101.1	[47]
PCFO71	n.d.	n.d.	n.d.	AY513487.1	[48]
CS5-like group					
CS5	H	5 nm	18.6	AJ224079	[49]
CS7	H	3-6 nm	18.7	AY009095.1	[50]
Class Ib group[3]					
CS12	F	7 nm	17.9	AY009096.1	[51]
CS18	F	7 nm	18.5	AF335469.1	[52]
CS20	F	7 nm	17.5	AF438155	[53]
CS26	n.d.	n.d.	n.d.	HQ203050.1	
CS27A	n.d.	n.d.	n.d.	HQ203047	[38]
CS27B	n.d.	n.d.	n.d.	HQ203047	[38]
CS28A	n.d.	n.d.	n.d.	HQ203049	[38]
CS28B	n.d.	n.d.	n.d.	HQ203046	[38]
CS30	F	7 nm	18.5	n.a.	Paper II
Additional					
CS3	f	2-3 nm	15.0	FN822745.1	[41,43,54]
CS6	nF		15.1/15.9	U04844	[45,55,56]
CS8	F	7 nm	25.3	AB059751	[57]
CS21	F	7 nm	25.2	EF59570.1	[58]
CS15	nF	n.d.	18.2	X65623	[59]
CS22	f	n.d.	15.03	AF145205.1	[60]
CS10	nF		n.a.	n.a.	[61]
CS11	f	3 nm	n.a.	n.a.	[62]
CS13[4]	f	n.d.	24.8	X71971	[63]
CS23[4]	f/nF	n.d.	16.9	JQ434477	[64]

[1]F = fimbrial; f = fibrillar; nF = non-fimbrial; H = helical. [2]The size of the major subunit was predicted using the published amino acid sequences. [3]All CFs in Class 1b are related to the porcine CF 987P [38]. [4]CS13 and CS23 are related to the porcine CF K88 [64]. n.d. = not determined. n.a. = not available.

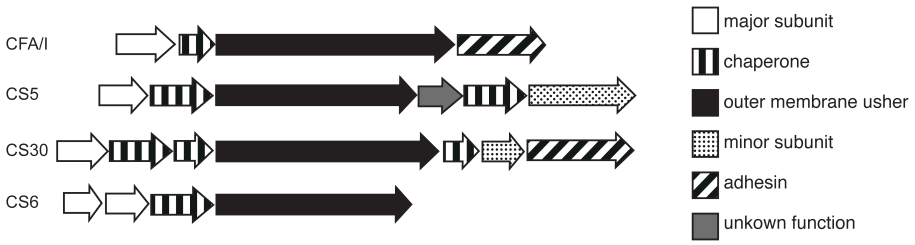


Fig 2. Operon structure of different ETEC CFs. Figure depicts the variation in operon structure between the ETEC CF groups. Representative CFs from the CFA/I-like group (CFA/I), CS5-like group (CS5) and Class 1b group (CS30), as well as the distinct CS6.

isolates. STp was originally identified in a porcine ETEC isolate and CFs of the Class 1b group are related to the porcine CF 987P [38](Paper II).

ETEC CFs may also be grouped based on conserved protein motifs. The usher and chaperone genes are the most conserved genes in CF/adhesin operons in Gram-negative bacteria. A clustering scheme for adhesins in Gram-negative bacteria based on specific protein motifs identified in the usher and chaperone(s), has been developed [28]. The classification of ETEC CFs using this scheme broadly corresponds to the ETEC CF-specific groups described above. CFA/I-like and CS5-like groups belong to the α-fimbrial clade and have similar operon structures. CS12 and CS18 belong to the Class 1b group and are assigned to the γ_2-fimbrial clade together with the porcine CF 987P. The porcine CFs, K88 and F18, are similar to the human ETEC CFs CS13 [14] and CS23 [64] and are located in the κ-fimbrial clade.

Distinct human ETEC CFs, CS3 and CS6, which express thin fibrillar and non-fimbrial structures, respectively, cluster together with adhesins of the AFA/Dr/AAF family found in UPEC, EAEC and DAEC. All CFs are plasmid-encoded, except CS2 which is encoded by chromosomal genes [65].

CF receptors/binding structures

Only a few of the known ETEC CFs have been studied in terms of receptor binding. CFA/I binds to glycoproteins and glycolipids expressed by epithelial cells of the human small intestine [66-68]. Furthermore, it has been shown that CFA/I binds to glycosphingolipids with the Lewis determinant, Le^a, through the major subunit, CfaB [68]. A subsequent study indeed showed that children with the Le(a+b-) blood group were more susceptible to diarrhea caused by bacteria expressing CFA/I [69]. It has been suggested that CS1, CS2, CS3 and CS4 bind

to the glycolipid asialo-GM_1 (lacking the sialic acid terminal) [70]. The non-fimbrial CF CS6 has been shown to bind to sulfatide through one of the major subunits (CssB) of CS6 [71].

Prevalence of ETEC CFs

The prevalence of specific CF profiles may vary depending on year and geographic location [5,72-75]. The most prevalent CFs are CFA/I and CS1-CS6, but other CFs like CS7, CS14 and CS17 have been shown to be common in ETEC in certain geographic regions. Additional CFs are CS8 (CFA/III), and CS21 (Longus) which also is considered a prevalent CF but may not be as important for virulence as other CFs [35]. CS8 and CS21 are classified as Type IV-like fimbriae and share a similar operon structure [76]. A putative subtype of CS8, CS8B has also been identified [77]. CS15 and CS22 are considered related but have a non-fimbrial and a fimbrial structure, respectively.

Individual ETEC isolates typically express one, two or three CFs, although some CFs are frequently co-expressed with specific CFs and/or toxin types: CS1+CS3 LT+STh, CS2+CS3 LT+STh, CS5+CS6 LT+STh, CS6 STp, CFA/I STh, and CS7 LT have repeatedly been isolated worldwide from patients with diarrhea [5,14,35,78].

Studies on the prevalence of different CFs in ETEC have shown that at least 30% of all clinical isolates do not express any of the CFs identified so far [12,35], suggesting that additional CFs may remain to be identified. Most ETEC CFs are plasmid encoded [14,79] and are frequently flanked by remains of insertion or transposase sequences [80] (Paper I), indicating that they are acquired by horizontal gene transfer.

Other bacterial adhesive structures

ETEC CFs are believed to be restricted to ETEC isolates, however additional operons encoding fimbriae or adhesins found in other pathogenic Gram-negative bacteria isolates may be present in ETEC. Type I fimbriae are known to be expressed by many members of *Enterobactericeae*, including commensal *E. coli*, and was present in the majority of the ETEC isolates analyzed in Paper I. Type I fimbriae are the main CF of UPEC, known to be the cause of at least 80% of all urinary tract infections. *E. coli*, such as UPEC, expressing the adherence fibrillar adhesin (AFA) or Dr adhesin are often identified in children and pregnant women with urinary tract infections [81]. Furthermore, DAEC,

which are thought to cause diarrhea in young children, also express both AFA and Dr adhesins [82]. However, a robust typing scheme for identifying DAEC has not yet been developed [82]. The distantly related aggregative adherence fimbriae (AAF) is found in EAEC and several studies report such fimbriae as a key factor in EAEC pathogenesis [83]. Several AFA/Dr/AAF-related operons were identified in some ETEC isolates, described in Paper III.

Other pathogens, such as *Salmonella enteritidis* and *Salmonella* Typhimurium express curli, encoded by the *csgABCDEFG* locus, which has also been shown to be expressed by commensal and pathogenic *E. coli* [84] including ETEC (data not shown). Long polar fimbriae *(lpf)* are expressed in various *Salmonella spp.* [85] and were found to be involved in the adhesive properties of EHEC isolates [86]. Furthermore, genes encoding Lpf are also harbored by ETEC isolates, mainly in the phylogenetic group B1 (data not shown).

The commensal *E. coli* K-12 harbors several chaperone-usher fimbrial operons that are functional and which could contribute to the ability of *E. coli* to adhere and colonize artificial and epithelial cell surfaces [86]. Several of these operons were identified in the majority of the screened ETEC isolates lacking a known CF (data not shown).

Putative novel virulence factors

The genomic content of ETEC is highly diverse and the search for a virulence factor present in all or the majority of ETEC isolates has led to the identification of several putative novel virulence factors. Both plasmid and chromosomally encoded genes [87], identified by phenotypic and genomic approaches, have been proposed as candidate antigens for future ETEC vaccines. The most prevalent putative virulence factors found so far are a metalloprotease that degrades mucins, YghJ [88] and a putative adhesin, EaeH (95% identity with FdeC [89]) [90]. YghJ conferred protection with high efficacy, and EaeH was partly protective in a mouse sepsis model [2]. However, both these putative virulence factors are also present in non-pathogenic *E. coli* [87,91]. The serine protease autotransporter EatA has been shown to be immunogenic and proposed as a vaccine antigen. EatA contributes to virulence by degrading MUC2, the main constituent of the mucus layer in the small intestine, allowing the bacteria to access the epithelial layer [92]. However, *eatA* was only found in 40% of the 362 ETEC isolates studied in Paper I [93]. A two-partner secretion locus *etpBAC* was identified by Fleckenstein *et al.* and they showed that the encoded

EtpA protein was located on the tip of the flagella, mediating binding to epithelial cells [87]. Furthermore, EtpA was shown to be protective against intestinal colonization of mice when vaccinated with EtpA glycoprotein [87]. Both *eatA* and *etpA* have been identified in ETEC of multiple phylogenetic backgrounds [94]. Other less prevalent putative virulence factors are CexE [95], Tib [96] and Tia [97].

Whole genome sequencing

High throughput sequencing such as whole genome sequencing (WGS) has allowed for detailed phylogenetic analysis of many important pathogens. Studies of their spread on a global and local scale, as well as the more recent evolution in response to pressure from therapeutics and the human immune response, are made possible by these analyses.

The first generation sequencing technology based on termination sequencing was developed by Fred Sanger and colleagues in the 1970s and was used to sequence the first set of pathogenic bacteria [98]. The first sequenced *E. coli* genome was published in 1997 [99] and was sequenced by shotgun sequencing technology [100]. In the late 2000s, next generation or high-throughput sequencing was developed, generating more data for a lower cost (Illumina and 454-sequencing). However, both Illumina and 454-sequencing generate short reads, making it difficult assemble a genome across repetitive segments. The first long-read technology was single-molecule real-time (SMRT) sequencing developed by Pacific Bioscience in 2010 [100]. This approach can generate high-quality assemblies and allows for detailed analysis of both chromosomes and plasmids.

Understanding bacterial evolution and transmission using genomics

Whole genome sequencing generates genomic data with high resolution, documenting single nucleotide polymorphisms (SNPs) across the genome, compared to traditional typing methods, which concentrate on SNP variations in one or several core genes such as multi locus sequence types (MLST). The high resolution of WGS compared to MLST, makes it possible to differentiate between closely related isolates in a clone. Also, the next-generation platforms of today can sequence multiple bacterial strains in one run, generating large amounts of data on a large collection of isolates, making it possible to study the

bacterial population structure at both a local and global level. Several studies have been published that show the strength of high-throughput sequencing to track the evolution and pathways of various pathogenic bacteria [101-106].

Using comparative genomics, several possible antigens have been identified in ExPEC isolates which are also present in intestinal *E. coli* [107]. Furthermore, Roy *et al*. identified 40 immunogenic antigens in the ETEC reference strain H10407, both known and putative virulence genes, as well as house-keeping genes [108], however the prevalence of these putative antigens across the ETEC population is unknown. It has also been proposed that a universal vaccine consisting of four antigens may protect against both pathogenic extraintestinal and intestinal *E. coli* [2]. However, the genes encoding these four proteins are also present in commensal *E. coli* [93]. Furthermore, in Paper I, II and III included in this thesis we have shown how using Next generation sequencing (Illumina) and Third generation sequencing (PacBio) can increase our knowledge of ETEC - not only the evolutionary patterns and the genomic content, but also how we can use WGS data to identify putative novel CFs.

ETEC genomics and genetic diversity

At the time when Paper I was published, the ETEC population structure and genomic composition were considered poorly characterized. Only two complete genomes (H10407 [109] and E23477A [110]) and six draft genomes had been sequenced. Where five of the draft genomes had been isolated in Guinea Bissau [94]. Now, 578 ETEC isolates have been sequenced, including the largest collection comprising 362 whole genome sequenced isolates described in Paper I.

Previous studies [94,109-111] have failed to identify genes that are exclusively present in the ETEC isolates examined and showed that ETEC is highly diverse. However, a study based on five whole genome sequenced ETEC isolates from Guinea-Bissau indicated that ETEC may have a conserved core genome [94]. Researchers have also implied that the mobile elements toxins and CFs determine an ETEC [94,112,113].

****

Whole genome sequencing is a powerful tool for studying the genomic content, evolution and transmission of pathogens. Hence, in this thesis we created the largest collection of whole genome sequenced ETEC strains to date. The genomic data, was used to study the population structure of ETEC, and to identify novel CFs that may be of importance in combating ETEC disease.

AIMS

The overall aim of this thesis was to explore the evolution of ETEC on global level utilizing high resolution whole genome sequence data from a large number of ETEC isolates and to identify novel putative colonization factors using reverse and forward genetics.

The specific aims were:

- To generate whole genome sequence data from a selection of 362 ETEC isolates collected during several decades from widely varied geographic regions.
    - To explore the population structure and evolution of ETEC based on whole genome sequencing.

- To identify novel putative colonization factors by two different approaches:
    - By reverse genetics: using the whole genome sequence data followed by phenotypic analyses.
    - By phenotypic analyses followed by genomic analyses.

MATERIALS AND METHODS

Bacterial isolates (Papers I-III)

In total, 362 human ETEC isolates from the University of Gothenburg ETEC strain collection, comprising more than 3,500 ETEC isolates, were selected based on major virulence factor profile (CF and toxin), origin and year of isolation, to represent a broad collection of ETEC isolated worldwide. This collection attempted to cover the most prevalent CFs as well as less common CFs and isolates lacking known CFs. The isolates selected were from Africa, Asia and the Americas and collected between 1980 and 2011.

ETEC isolates from all patient groups were included, i.e. children below five years of age and adults from ETEC endemic areas, as well as travelers and soldiers visiting such areas. The majority of the isolates have been collected from individuals suffering from diarrhea, both hospitalized and outpatients, and some were derived from asymptomatic carriers. All ETEC isolates were identified by culturing on MacConkey agar followed by LT and ST toxin expression analysis by ELISA and in many cases toxin PCR [114]. The colonization factor profile was determined by dot-blot analyses [114] and in some cases by multiplex PCR (Paper I) [115]. For Paper II and III a subset of the 362 ETEC isolates were used, specifically isolates lacking a known CF ("CF negative").

Growth conditions (Paper I-III)

For DNA preparation, the 362 ETEC isolates were cultured overnight on CFA agar plates containing crude bile [116] as described in Paper I. The "CF negative" isolates were cultured overnight on CFA agar plates containing crude bile for SDS-PAGE. The same cultures were used for subsequent culturing in CFA broth or on agar plates containing crude bile for Caco-2 adhesion assay and transmission electron microscopy (TEM) imaging as described in Paper II and III.

Whole genome sequencing (Papers I-III)

Next generation sequencing and de novo assembly

In total, 362 whole genome sequenced ETEC isolates were used in Paper I-III. A selection of 374 isolates were sequenced by Next Generation sequencing on a HiSeq Illumina platform generating reads either 75 or 100 bases long. These reads were used for de novo assembly using the Velvet assembler [117] generating contigs. Successful assemblies, based on average bases assembled, average number of contigs and N50 (a weighted median, 50% of the entire assembly is contained in contigs of equal or larger than this value). Samples that were contaminated or failed to be properly assembled were discarded or re-sequenced. In total, 362 ETEC isolates passed the assembly quality check and used for further genomic analyses (**Fig 3**). Since the whole genome sequences were planned to be used for searching for novel CF genes/operons, de novo assembly was chosen since alignment of whole genome sequenced isolates to a reference genome does not include non-reference sequences (Paper I).

Identification of ETEC core genes

Determination of the core genes, i.e. genes that are chromosomally encoded and present in all isolates examined, were used for studying the population structure. The maximum common genome (MCG) was determined by analyzing 47 fully sequenced and annotated *E. coli* strains available in Genbank (June 2012), including the ETEC reference strain H10407. Genes present in all 47 *E. coli* strains with at least 70% identity on the protein level were identified using hierarchical clustering by CD-HIT and considered part of the MCG. The 1,429 core genes identified were extracted from the 362 ETEC isolates and used for phylogenetic analyses (Paper I) (**Fig 3**).

Phylogenetic analyses and identification of lineages

Although homologous recombination in *E. coli* is generally low [118], it may interfere with the population structure. Therefore, significant recombination regions were masked using the BratNextGen method [119] on the concatenated MCG alignment. The phylogenetic tree was based on the SNP alignment with recombination sites masked and generated by FastTree (v2.1.4). Twenty-one well-characterized *E. coli* references were included in the phylogenetic analysis

(Paper I). To estimate the population structure, we applied Bayesian analysis on the population structure (BAPS) [120] which uses nested clustering to fit lineages to genome data. The analysis was performed on the MCG alignment with recombinations masked as missing data (**Fig 3**). In total, 21 lineages were identified across the ETEC phylogeny.

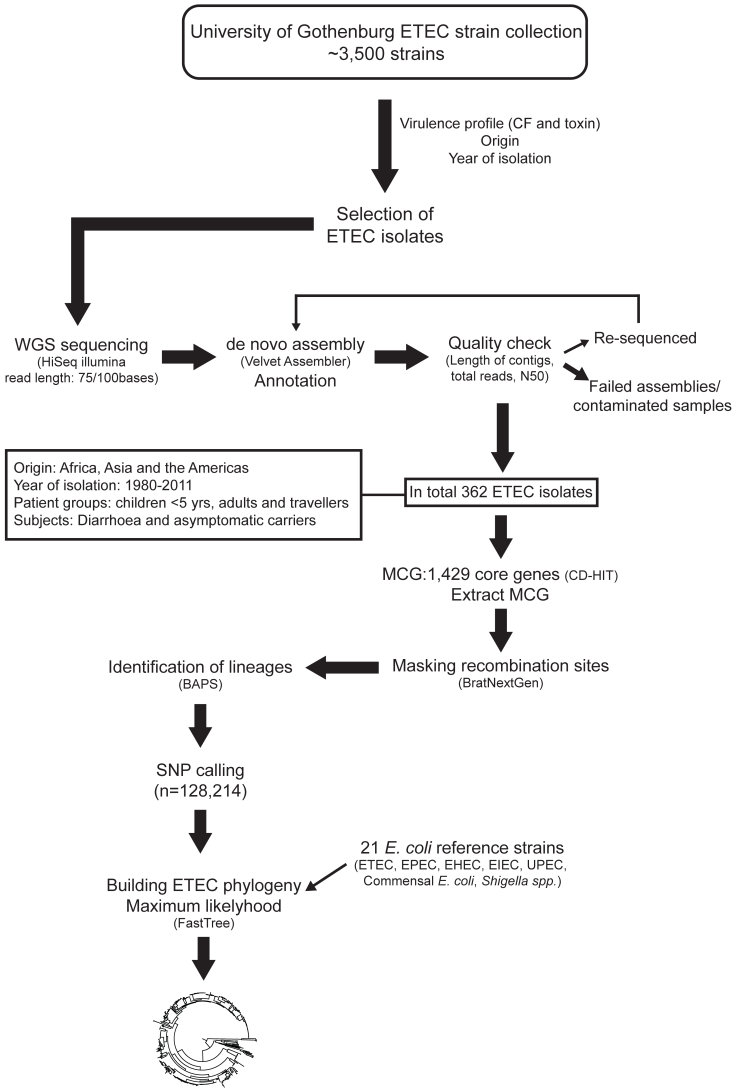


Fig 3. Flowchart illustrating the generation of the ETEC phylogeny based on 362 selected ETEC isolates. The isolates are collected from children, adults and travelers with diarrhea, and some asymptomatic carriers, from Africa, Asia and the Americas between 1980 and 2011.

Using comparative genomics to identify putative novel CFs

When searching for putative novel genes/operons, in this case encoding putative novel CFs, different approaches are needed. For this purpose we have used a sequence comparison tool (BLASTn) and a protein motif comparison tool (HMMER).

BLAST: regular sequence comparison (Paper II-III)

Published sequences of operons encoding ETEC CFs and additional adhesins from other Gram-negative bacteria were concatenated and used as a reference database for a BLASTn search to identify sequences that may encode a putative CF.

HMMER: searching for conserved usher protein motifs (Paper III)

Ushers of fimbriae operons are known to be conserved [28]. A specific protein motif has been identified in the fimbriae part of the alternate and classical CU pathway, PF00577. Hence, a database of protein sequences encoding all ETEC ushers containing the protein motif PF05577 was created. This database was used to perform a HMMER iterative search [121] for related proteins in the open reading frames (ORFs) of the whole genome sequences of the "CF negative" isolates.

Roary: Prokaryotic pan genome analysis (Paper III)

To ensure that the identified sequences that may encode putative CFs are not present in the ETEC isolates harboring known CFs or in the majority of isolates, the pan genome of all 362 ETEC isolates, from Paper I, was determined using Roary. This tool combines results from two different clustering tools, MCL and CD-HIT [122]. Roary was run with default settings on the Sanger servers.

Phenotypic analyses

Toxin production (Paper II-III)

The toxin production was determined by LT GM1 ELISA and ST inhibition GM1 ELISA as previously described [114].

SDS-PAGE (Papers II-III)

ETEC CFs are known to be thermo-regulated [14,123,124], hence heat-extracts from cultured ETEC isolates at either 20°C or 37°C were analyzed by SDS-PAGE, as described before [52], to identify thermo-regulated proteins of a size between 12 and 25 kD, a size range typical for major subunits of ETEC CFs [14].

Adhesion assay (Papers II-III)

The small intestine-like cell line Caco-2 was used for the adhesion assay. Caco-2 cells were seeded in 8-well chamber glass-slides and grown for 14 days post-confluence for differentiation, as described [52]. Bacterial cell suspensions were added to the differentiated Caco-2 cells and incubated for 3 hours followed by Giemsa staining and analysis in a light microscope. Isolates with three or more bacteria per Caco-2 cell after culture at 37°C compared to at 20°C were considered positive for binding (Paper III).

Transmission electron microscopy (Papers II-III)

Transmission electron microscopy (TEM) was used to examine whether "CF negative" isolates expressed outer membrane appendages, such as fimbriae or fibrillae. Bacterial isolates were cultured as described in Paper II and III and applied onto freshly glow-discharged Formvar-copper coated grids followed by fixation with glutaraldehyd and negative staining using phosphotungstic acid. The samples were examined and photographed in a Zeiss transmission electron microscope using a Veleta camera.

qRT-PCR (Papers II)

A growth curve was conducted to evaluate the time-point for maximum expression of CS30. Samples were taken every hour during seven hours and from the overnight culture. Total RNA was prepared and cDNA was synthesized by regular PCR. The expression of CS30 was measured by running quantitative real-time PCR (qRT-PCR) using primers specific for the major subunit *csmA*. The levels of transcripts were normalized against the sample collected after two hours of culture.

Mass spectrometry (Paper II-III)

In Paper II, quantitative mass spectrometric analysis of whole bacterial pellets from two different culture conditions, 20°C and 37°C, were used to evaluate the difference in production of CS30-related genes. As mentioned above, SDS-PAGE was employed to identify ETEC isolates with thermo-regulated proteins of a predefined size as described in Paper II. To ensure that the protein bands that were identified as the putative major subunit indeed represented a CF protein, regular mass spectrometry analysis was conducted on the ~18 kD band in all four CS30 positive isolates and on the ~15 kD band in E1526, part of lineage 16, harboring the AFA/AAF-related operons.

RESULTS AND DISCUSSION

The evolution and diversity of ETEC (Paper I)

Although ETEC is one of the most common causes of diarrhea, the genomic characterization of ETEC has not been examined in greater detail. At the time of this study, only two complete ETEC genomes and six draft genomes were available. Whole genome sequencing is a powerful tool for not only studying the evolution and population structure of bacteria, but also for examining the prevalence of immunogenic antigens and for identifying and characterizing potential novel vaccine targets. In Paper I, we created the largest collection of whole genome sequenced ETEC isolates to date. In total, 362 ETEC isolates were selected from the University of Gothenburg ETEC strain collection comprising >3,500 ETEC isolates. The isolates chosen were collected from various geographical locations between 1980 and 2011 from different patient groups: children below five years of age, adults as well as travelers and soldiers visiting endemic areas.

Population structure of ETEC

Studies on the ETEC population structure have mainly been based on either fingerprinting analyses of intergenic and repetitive repeats or by determining the multi locus sequence type (MLST) profile, which is done by studying the sequence variation in multiple chromosomally encoded genes (usually six to eight) [125]. Difficulties differentiating between closely related isolates limits these methods. ETEC has until now been considered as a highly diverse pathovar without a clear population structure, as ETEC are present across the *E. coli* phylogeny. WGS data can provide an in-depth comprehensive overview of the gene repertoire and sequence variation. Using the WGS data of the 362 ETEC isolates, we aimed to study the population structure of ETEC in association with O antigen, CFs and toxins.

Genetic material may be acquired by homologous recombination either by conjugation, transduction or transformation. Removing identified recombination sites reduces the risk of a false impression of the relationships between isolates. Building a phylogenetic tree based on single polymorphisms and with recombination events removed revealed distinct ETEC lineages that

were found across the *E. coli* phylogeny. Collectively, this suggests that ETEC isolates have arisen from several commensal *E. coli* clades. Thus, as previous studies have shown, ETEC falls into *E. coli* phylogenetic groups [94,113], mainly assigned to A and B1, which is the main phylogenetic groups for gut commensal bacteria [126].

Characteristics of ETEC lineages

The major virulence profile (CFs and toxins), origin and year of isolation were known for all whole genome sequenced isolates and the O antigen profile was determined using a novel *in silico* method (Paper I). Two thirds of the isolates were CF positive, i.e. they expressed known CFs as determined by dot blot analyses and in some cases also by multiplex PCR. The isolates that had previously been classified as lacking a known CF, referred to as "CF negative" in this thesis, were re-confirmed to lack any known CF by genomic analyses.

Linking the virulence factor profile with the high-resolution population structure in a phylogenetic tree indicated the presence of 21 lineages (L1-L21); isolates within the majority of these lineages shared specific profiles with regard to O antigen, CF- and toxin profile (L1-L11) (**Fig 4A**). Isolates found in the major lineages (L1-L5) expressed the most prevalent ETEC CFs described in the literature [12,14]: CS1+CS3 (L1), CS2+CS3 (L2), CFA/I (L3), CS6 (L4) and CS5+CS6 (L5). The isolates within each major lineage also shared a specific toxin profile (**Fig 4B**). These results indicate that the virulence factors are ancestral and that the distinct lineages have spread by clonal expansion. This assumption is at variance with those of previous studies stating that the CF and toxin genes are randomly distributed and that the acquisition of virulence genes most likely has occurred at multiple times through HGT [94,111-113]. A recent study on the ETEC population structure during infection [127] showed that clones from patients were closely related to temporally and geographically dispersed strains from other diarrhea cases. This finding indicates that, despite the genetic diversity, virulence traits may be maintained over time and also suggests that the core genome and acquisition of virulence factors work in concert. This is in line with the findings in Paper I.

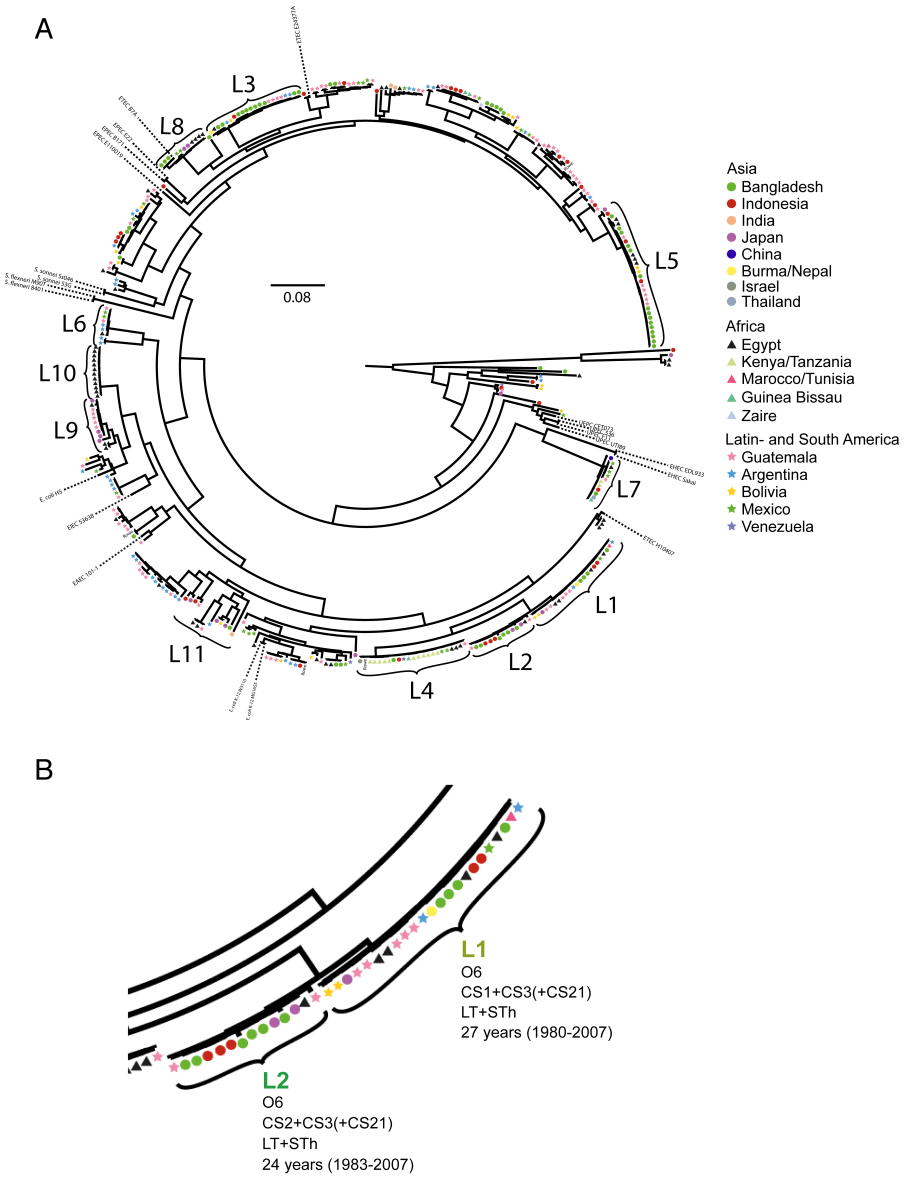


Fig 4. Phylogenetic tree showing the population structure of ETEC. A. Midpoint rooted circular phylogenetic tree of ETEC based on the sequence variation detected in the maximum common genome (MCG) with recombination sites masked. Lineages 1-11 (L1-L11) are indicated, as well as *E. coli* references. The origin of each isolate is shown by colored symbols on the tips of each branch. Scale bar: 0.08 substitutions per variable site. **B.** Enlarged section showing lineage 1 and lineage 2 with countries specified on branch tips. Virulence profile and year of isolation are indicated.

A variation in the O antigen profile was observed within the identified lineages. For example, in L3 four subgroups with different O antigens were identified and the largest lineage, L5, could be divided into two subgroups: one with O115-positive isolates expressing either CS5+CS6 or CS17 and the other with O167 expressing isolates expressing CS5+CS6. However, most of the lineages with a highly variable O antigen profile contained "CF negative" isolates or isolates expressing less prevalent CFs. In addition, although the "CF negative" isolates are more diverse, the four identified lineages (L11-L14) predominantly harboring "CF negative" isolates exhibit a phylogenetic structure, indicating that they may share genomic properties such as not yet identified novel CFs.

To study the transmission of pathogen isolates, we need to know where the isolates had been collected and at what time-point they had been isolated. Coupling this information together with the high-resolution genotype of each ETEC isolate revealed that some lineages are globally and temporally distributed, indicating a clonal spread. This finding of a close clustering of isolates from different parts of the world collected over a period of 30 years suggests that the acquisition of plasmid-encoded virulence factors occurred once and was then followed by a clonal expansion of isolates carrying the same virulence factors.

As whole genome sequencing has become more robust and cheaper, several genome-based studies on a global level of different pathogens have been published, similar to the studies described in Paper I. Lineages identified for other pathogens, such as *Shigella dysentarie* and *Shigella flexneri* have geographic affinities which we did not find for ETEC [102,128,129]. The reason for this discrepancy is most likely due to the enormous diversity in ETEC.

Not only the acquisition of virulence genes may be important for maintaining virulence traits, but also loss of specific genes, so called reductive evolution. Loss-of-function gene mutations result from bacterial adaptation to a more specific niche, where superfluous genes are lost, deleted or differentially expressed. Virulence is of key importance for host-restricted pathogens, thus genes that interfere with virulence may be selected against; such genes are called anti-virulence genes (AVGs) [130].

Estimating the emergence of major ETEC lineages

In an attempt to estimate the emergence of the major ETEC lineages, a Bayesian modeling technique, BEAST, was used. BEAST links the accumulated nucleotide variations to the year of isolation, moving back in time to reconstruct the evolutionary history allowing estimation of the most recent common ancestor (MRCA). Such analysis showed that lineages 1 through lineage 5 emerged between 51 and 174 years ago (**Table 2**). CFA/I and CS6-specific lineages were estimated to be the oldest lineages. Isolates with these CF profiles have been identified in additional locations in the ETEC phylogeny tree. These results indicate that the lineages are temporally stable and imply a tighter and long-term association between the chromosome and plasmids than has previously been appreciated. This observation is coherent with the spread of other pathogens, such as *Vibrio cholera* [106], *Staphylococcus aureus* [101], *Shigella sonnei* [105] and invasive *Salmonella* Typhimurium [102]. The substitution rate, i.e. the rate of mutations becoming fixed in the genome, was largely consistent in the major ETEC lineages, and similar to the substitution rates observed for *Streptococcus pneumonia* [131] and *Clostridium difficile* [132]. The estimated number of SNPs accumulated each year was 2-5.5 (depending on the lineage) (**Table 2**), which is similar to the number of SNPs accumulated in *Vibrio cholera* (3.3 SNPs per year) [106].

Table 2: Characteristics of the major ETEC lineages L1-L5.

Lineage	No. of isolates	MLST	O antigen	CF	Toxin profile	Time of emergence[1]	Substitution rate
L1	23	ST2353[2]	O6	CS1+CS3[3]	LT+STh	1955	1.0×10^{-6}
L2	14	ST4	O6	CS2+CS3[3]	LT+STh	1961	1.0×10^{-6}
L3	22	ST173	O78, O114, O126, O128	CFA/I, CS7	LT+STh	1837	3.7×10^{-7}
L4	23	ST1312	O25	CS6[3], CS6+CS8, CS21	LT, STh	1874	4.0×10^{-7}
L5	30	ST443	O115, O157	CS5+CS6, CS17	LT+STh, LT, STh	1930	1.1×10^{-6}

[1]The median values from BEAST analysis. [2]Four isolates belonged to ST4. [3]With or without CS21 (Longus).

Collectively, these results reveal persistent plasmid-chromosomal background combinations, which have been stable over substantial periods of time in endemic areas. Our data suggest that plasmid acquisition is a major event driving the emergence of different ETEC lineages. Hence, it is likely that an ETEC vaccine based on the most prevalent CFs [12] could protect against a large proportion of ETEC diarrhea cases. However, at least one third of all clinical ETEC isolates have been characterized as "CF negative" in different studies and such ETEC isolates will most likely not be targeted by a future ETEC vaccine based on already identified CFs. Hence, it is of great importance to search for additional prevalent ETEC CFs that may be included in future ETEC candidate vaccines to increase protective coverage.

Using two different approaches to search for putative novel ETEC CFs (Paper II and Paper III)

CFs mediate binding of ETEC to epithelial cells in the small intestine. Adherence is a key part of ETEC pathogenesis, hence ETEC isolates known to cause diarrhea are assumed to carry genes for expressing CFs. "CF negative" isolates may either colonize the intestine through as of yet not identified CFs or have lost CF encoding genes at the time of analysis [12,35,133]. We have designed two different approaches for identifying putative novel CFs. In Paper II, we used reverse genetics, starting with a comparative genomic analysis of 94 whole genome sequenced "CF negative" isolates followed by phenotypic analyses: SDS-PAGE, adhesion assays and electron microscopy and subsequently third generation sequencing (**Fig 5**). In Paper III, we started by screening 85 "CF negative" isolates for thermo-regulated proteins of a predefined size using SDS-PAGE, followed by analyses of the isolates in an adhesion assay, transmission electron microscopy and finally comparative genomics, to identify putative ETEC CF loci (**Fig 5**).

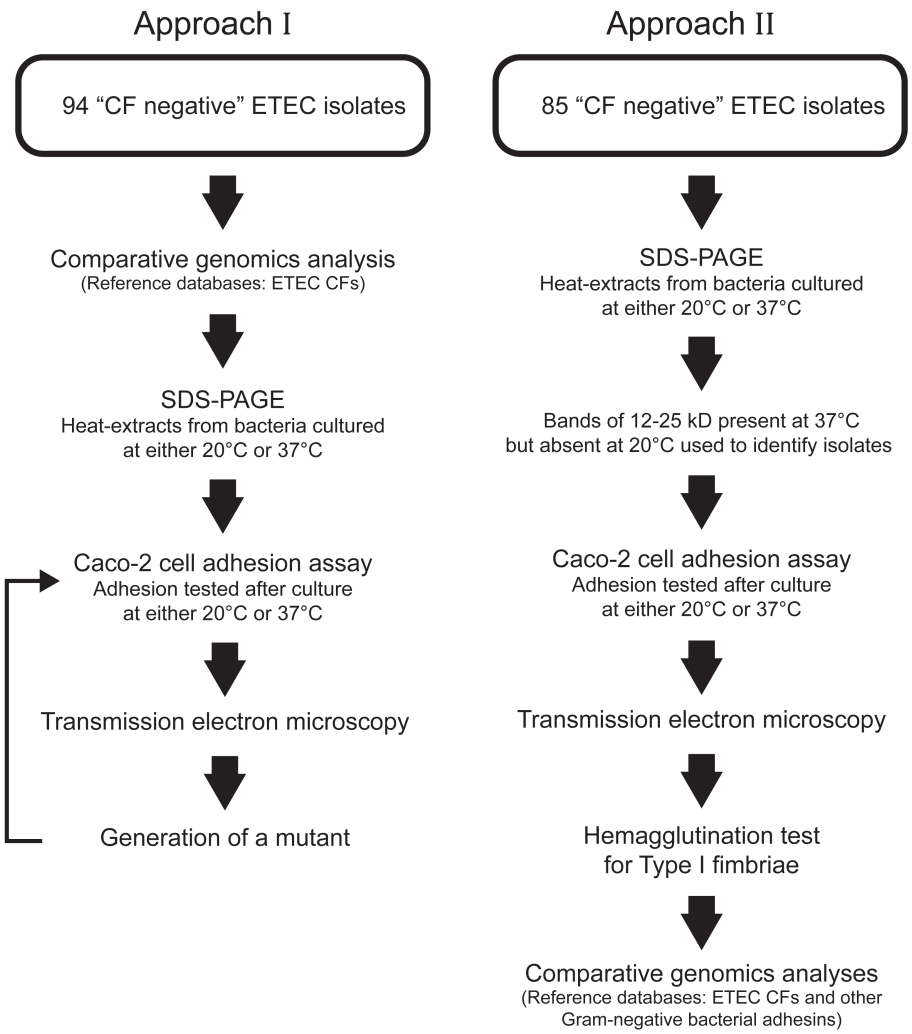


Fig 5. Flowchart depicting the workflow for identifying and characterizing putative novel ETEC CFs according to approach I and II.

Identification of CS30 (Paper II)

Searching for novel CFs by reverse genetics

From a selection of 94 whole genome sequenced “CF negative” isolates, a comparative genomics approach was used by mapping reads of the “CF negative“ isolates against a reference database of all known ETEC CFs. Six isolates were found to carry a loci of seven genes organized as an operon, two of

which also harbored genes for the porcine-like CF CS13 [14]. The operon structure was found to be very similar to that of the human CF CS18. CS18 fimbriae are both antigenically and structurally similar to CS12, CS20 and the porcine CF 987P (F6) [14,134]. The six "CF negative" isolates identified had been collected from children below five years of age with diarrhea from Argentina, Egypt and Guatemala between 1989 and 2003; they were located across the ETEC phylogenetic tree (Paper I). The mature major subunit encoded by the operon was predicted to be 18.5 kD, identical to that of CS18. The highest amino acid sequence similarity of this protein was found with the major subunit CsnA of CS20, FotA of CS18 and FasA of 987P. In addition, two site-specific recombinases were identified upstream of the putative major subunit, similar to CS18. A phylogenetic tree based on the amino acid sequences encoding the major subunit of known ETEC CFs (available CFs in GenBank) showed that the putative major subunit of the four "CF negative" isolates lacking CS13 formed a distinct cluster together with other CFs of the Class 1b group. Ushers are considered having conserved amino acid sequences with specific protein domains identified [28]. Hence, a protein analysis of the sequence encoding the putative ushers of the four isolates using the Pfam database revealed the well known protein motif PF00577, present in all ushers part of CFs assembled by the CU pathway [28].

The genomic analyses indicate that the loci encodes a novel CF. We have chosen to designate this novel CF CS30 and the seven genes of the operon *csmA-csmG* with two site-specific recombinases, *csmS* and *csmT*, located upstream of the operon, according to the nomenclature introduced by Gaastra and Svennerholm [14].

Phenotypic analyses of isolates harboring CS30

Virulence factors in ETEC isolates are usually regulated by external signals from the environment, such as temperature [14,64,123], pH [135], bile [116] and glucose [136]. We examined whether the four CS30 positive "CF negative" isolates produced thermo-regulated proteins with a size between 12 and 25 kD, which corresponds to the size of a major subunit, and we could show that all the isolates expressed polypeptides of ~18 kD in size after culture at 37°C but not at 20°C.

The next step was to examine the phenotypic traits of the four isolates. The well-characterized small intestine-like cell line Caco-2 is widely used to study adherence of bacteria. Using differentiated Caco-2 cells, we showed that the CS30 positive isolates bound well after being cultured at 37°C but not at 20°C. Furthermore, knowing that CS30 is related to CS18 and CS20, both producing fimbrial structures, transmission electron microscopy (TEM) imaging of a representative CS30 positive isolate (E873) was performed showing heavily fimbriated bacteria (**Fig 6**).

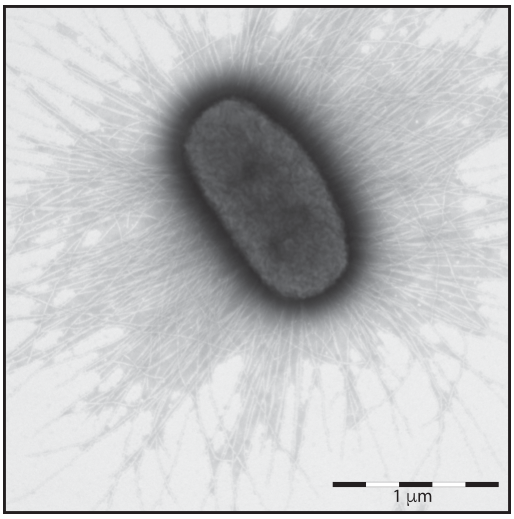


Fig 6. Transmission electron microscopy image of isolate E873. Showing heavily fimbriated bacteria by phosphotungstic staining.

Most ETEC CFs have a fimbrial structure built of homopolymers of a major subunit, with a minor subunit on the tip of the fimbriae that may act as an adhesin [14]. To show that the identified CS30 operon was producing the fimbriae visualized by TEM, a mutant with a disrupted gene encoding for the major structural subunit was constructed. TEM imaging of the mutant showed no fimbrial structures.

CS30 is part of the Class 1b group known to be related to the porcine CF 987P [38]. In a study where 1,019 human ETEC isolates together with 8 porcine ETEC isolates were phylogentically analyzed (based on MLST data), it was shown that six of the eight porcine ETEC isolates clustered together with three different human ETEC lineages, suggesting that human and porcine ETEC have a common ancestor [111]. Our findings show that the human CF CS30 not only is related to the porcine CF 987P but CS30-positive isolates also

express STp, which is produced by porcine-specific ETEC isolates. This indicates that porcine CFs may have been acquired through horizontal gene transfer or have evolved from a common ancestor. Other human ETEC CFs related to porcine CF are CS13 [14] and CS23 [64].

Isolates harboring CS30 have different chromosomal backgrounds; two isolates cluster together with a CS18 positive isolate and seem to be related to *Shigella sonnei* Ss046 and 536 in the ETEC phylogenetic tree. The third isolate clusters together with two CS13+CS30 positive isolates and the forth isolate is located on a distinct chromosomal background with the closest relative being two *E. coli* K-12 isolates. One could speculate that CS30 evolved a long time ago, and can be retained and stable on multiple core backgrounds. Similarly to the observations for CFA/I and CS6 positive isolates that are part of several distantly related lineages.

PacBio sequencing of E873

In 2003, the third generation sequencing technology was developed, Single Molecule Real Time or PacBio sequencing, allowing longer segments of DNA to be sequenced [137]. This technique allows for sequencing and circularization of both chromosome and eventual plasmids that can be further studied. Genomic analyses of the CS30 positive isolates indicated that the CS30 operon was plasmid-encoded, both based on the short contigs they were identified on and the multiple insertion sequences and transposon-related genes flanking the operon. Using PacBio sequencing and Circlator [138] allowed us to close the chromosome and circularize three plasmids, E873p1-p3, named according to size. E873p1 harbored antibiotic resistance genes along with replication and transfer-related genes, while E873p2 did not contain any antibiotic resistance or virulence genes. E873p3, the smallest plasmid, harbored an intact CS30 operon, including two site-specific recombinases upstream of the major subunit, and *eltA* and *eltB* encoding LT and *estA* encoding STp. In addition to the virulence genes, replication and some transfer-related genes were present on this plasmid.

Mobilization of a plasmid lacking transfer genes through the aid from another plasmid has been shown to work in several pathogens [139-142]. Hence, the presence of an origin of transfer *(oriT)* site, which is a short sequence needed for the transfer of DNA during conjugation, suggests that the plasmid probably is mobilizable with the help of the transfer genes in E873p1.

Using phenotypic analyses followed by comparative genomics to search for novel putative CFs (Paper III)

Phenotypic analyses of "CF negative" isolates

The majority of ETEC CFs have been shown to be thermo-regulated [52,64]. As mentioned before, the main building blocks of ETEC CFs are the major subunits that are usually between 12 to 25 kD in size. Therefore, heat-extracts of 84 "CF negative" isolates cultured at 37°C and at 20°C were analyzed by SDS-PAGE and screened for thermo-regulated proteins of 12-25 kD in size. Altogether, 35 isolates expressed polypeptides at 37°C which were absent after culture at 20°C and these isolates were further evaluated for putative novel ETEC CFs. Analyses of the 35 isolates after culture at 37°C versus at 20°C for binding in the Caco-2 cell assay reduced the number of isolates harboring potential novel CFs to nine.

The nine isolates adhering well to Caco-2 cells were screened for visible protruding structures by TEM. Two of these isolates had visible fimbrial structures after culture at 37°C which was absent after culture at 20°C. However, one of the isolates (E1360) was shown to be positive for Type I fimbriae, hence we cannot exclude that the visible fimbriae were Type I fimbriae due to lack of a specific immunoreagent for such fimbriae. The other isolate (E5087) only had a few visible fimbriae present on some bacteria but was negative for Type I fimbriae.

Genotypic analysis of nine "CF negative" isolates

The WGS data (Paper I) of the "CF negative" isolates analyzed in the studies described in Paper III were used for comparative genomic analyses to search for putative CF operons. Three different approaches were used: BLASTn analysis using two reference databases to identify similar gene/operons to already known CFs and/or adhesins, protein motif analysis (HMMER) to identify conserved usher proteins and a clustering method (Roary) to identify isolates with unique genes.

The two different BLASTn reference databases included all available ETEC CFs (whole operon or major subunit) and adhesins found in other Gram-negative bacteria, respectively. The BLAST and HMMER analyses revealed four putative CF operons related to the AFA/Dr/AAF family of

adhesins, found in UPEC, EAEC and DAEC [82,83,143-145]. Two putative CF operons, the AFA-like operon I and the AAF-like operon I, were identified in five closely related isolates (E1526, E1623, E1637, E1674 and E2348) (**Table 3**), part of lineage 16 (L16). These isolates had been collected from children with diarrhea in Argentina, Bolivia and Indonesia between 1989 and 2007. Mass spectrometric analysis of expressed proteins in the 15 kD band of isolate E1526 revealed peptides matching the adhesin AfaE part of the AFA-like operon I. This finding indicates that the AFA-like I operon may encode for the adhesin mediating adhesion to Caco-2 cells. Furthermore, AFA adhesins have shown to bind to the glycoconjugates human decay accelerating factor (hDAF), also called CD55, and the carcinoembryonic antigen-related cell adhesion molecules (CEACAMs) [82], which are expressed by Caco-2 cells [146,147].

Isolate E5087, which is not part of any of the identified lineages and a distant relative to the L16 isolates, harbored two similar putative CF operons (AFA-like operon II and AAF-like operon II) to the ones found in the five L16 isolates (**Table 3**). The majority of the proteins encoded by these two operons (AFA-like operon II and AAF-like operon II) are highly similar to the AFA-like operon I and AAF-like operon I present in the five L16 isolates. However, the subunits known to be responsible for the adhesion are less similar. The adhesin subunits B encoded by the AFA-like operon I (L16) and the AFA-like operon II (E5087) are only 32% identical and the major subunits encoded by the AAF-like operon I (L16) and AAF-like operon II (E5087) share 75% of their amino acid sequences. Hence, the operons identified in the five L16 isolates and E5087 should be considered as four separate operons.

Furthermore, the BLASTn analysis identified another putative CF operon in E1360 related to the human CFs CS13 and CS23 and the porcine CF K88 (F4) (**Fig 7** and **Table 3**). Both CS13 and CS23 have previously been shown to be related to the porcine CF K88 (F4) [14,64].

Table 3: Phenotypic characteristics and genomic findings of Caco-2 cell binding ETEC isolates (Paper III)

Isolate[1]	Lineage	O antigen	Toxin profile	SDS-PAGE[2]	Caco-2 adherence[3]		EM[4]	Genomic findings[5]
					20°C	37°C		
E1526	16	ON5	ST	15	-	++	Neg	2 AFA/Dr/AAF-like
E1623	16	ON5	ST	15	-	++	Neg	2 AFA/Dr/AAF-like
E1637	16	ON5	ST	15	-	++	Neg	2 AFA/Dr/AAF-like
E1674	16	ON5	ST	15	-	++	Neg	2 AFA/Dr/AAF-like
E2348	16	ON5	ST	15	-	++	Neg	2 AFA/Dr/AAF-like
E5087	None	O102	ST	17	-	+	F	2 AFA/Dr/AAF-like
E1360	19	O21	ST	12 + 24	-	++++	F	K88/CS23-like

[1]All isolates were collected from patients with diarrhea and all except E1360 and E5087 were from children below 5 years of age. [2]Size (kilodalton) of protein band expressed after culture at 37°C but not at 20°C. [3]– = (< 3); + = (3-13); ++ = (14-24); +++ = (25-35); ++++ = (36-46) number of bacteria per Caco-2 cell. [4]F=Fimbrial. [5]Operons identified.

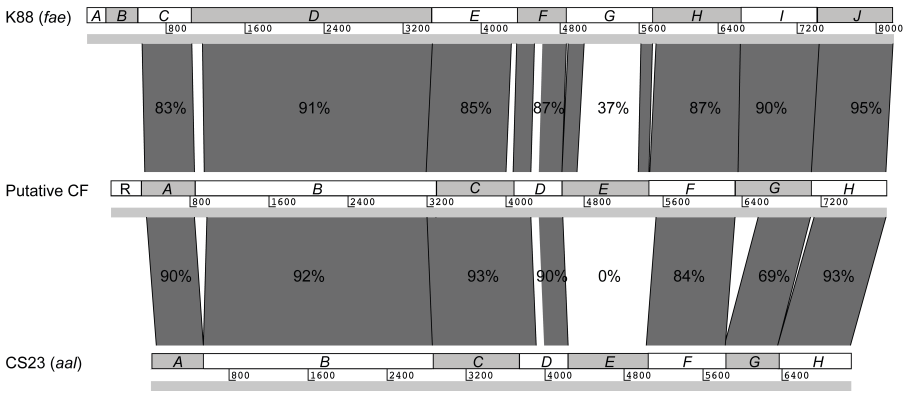


Fig 7. Comparison between the operons of K88 *(fae)*, putative CF (E1360) and CS23 *(aal)*. The extracted nucleotide sequences of the porcine ETEC CF K88 (F4) and the human ETEC CFs CS23 were compared to the identified putative CF operon in E1360. Percentages within gray blocks indicate the identity between the respective genes. K88 is encoded by the genes *faeA-faeJ* and CS23 is encoded by the genes *aalA-aalH*. Genes of the putative CF operon are designated *A-H* with R encoding a putative regulator.

Although the overall operon structure was highly conserved in K88, CS23 and the putative CF operon (**Fig 7**), the gene encoding the putative major subunit/adhesin *(E)* was shown to be distantly related to the major subunits/adhesins *faeG* and *aalE* of K88 and CS23, respectively. Furthermore, the putative regulatory gene in the putative CF operon (R) does not match the PapB-like regulator *(aalR)* in CS23 [64] or the regulators *(faeA and faeB)* in K88 [148], respectively. However, it is highly similar to *afaA*, the regulator gene of AFA-I found in UPEC and DAEC strains [82,145].

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

In summary, this thesis has shed light on the population structure and evolution of ETEC by using whole genome sequencing (WGS) data from a broad selection of ETEC isolates collected worldwide during three decades from children, adults and travelers with diarrhea. Building a phylogenetic tree based on core genes revealed several distinct ETEC lineages (L1-L21). The majority of lineages, comprising isolates with specific virulence profiles (O antigen, CFs and toxins), were globally and temporally distributed. The major lineages (L1-L5), comprising isolates with the most prevalent CF and toxin profiles, were estimated to have emerged during the last 170 years and seem to be stable over time. Previous studies of ETEC evolution have been MLST-based, i.e. using MLST for phylogenetic analyses, which is only based on six to eight core genes and does not allow to differentiate between closely related isolates. Due to the enormous diversity seen in ETEC, the lineages described in Paper I have not previously been identified and ETEC isolates have been considered to be determined by the random acquisition of virulence genes through horizontal gene transfer [94,111,112].

The results presented in Paper I revealed the presence of an earlier unknown association between the plasmid-encoded virulence genes and the chromosomal background, as well as that the acquisition of both CFs and toxins has occurred once and then spread by clonal expansion. In addition, ETEC isolates with the same, stable virulence profiles seem to cause disease in children, adults and travelers. Altogether, these results suggest that the development of a vaccine based on the most prevalent CFs could be protective against a large proportion of ETEC diarrhea cases. The collection of whole genome sequenced isolates will provide a sound basis for future studies on ETEC transmission and pathogenicity.

Although the majority of clinical ETEC isolates express known CFs, depending on the geographic location, at least 30% of all clinical isolates lack an identifiable CF [12,35]. By using various genomic tools and phenotypic analyses, two different approaches to identify novel ETEC CFs were developed. These approaches have resulted in the identification and characterization of a

novel CF (Paper II) and the identification of several putative CF operons (Paper III).

The first approach, described in Paper II, was based on reverse genetics, screening the whole genome sequences of "CF negative" isolates from Paper I for putative CF operons followed by phenotypic analyses, such as SDS-PAGE, adhesion assays and transmission electron microscopy (TEM). Using this approach, the novel CF, CS30, was identified, characterized and shown to have a fimbrial structure. The expression of fimbriae was thermo-regulated and when cultured at 37°C, bacteria expressing CS30 bound to Caco-2 cells. The major subunit of CS30, CsmA, was shown to be 18.5 kD in size and when the gene encoding CsmA was disrupted no fimbriae were expressed. The next-generation sequencing used in Paper I generates short reads, which were de novo assembled. Highly repetitive segments, present in plasmids, are therefore difficult to assemble correctly. When using third generation sequencing, PacBio, it was possible to identify and circularize the plasmid harboring CS30, which also carried the genes encoding LT and STp. CS30 was found to be related to the porcine CF 987P, which may suggest that host-specific ETEC CFs may be reciprocally transferred between humans and animals, e.g. pigs.

The second approach, described in Paper III, was based on initial SDS-PAGE analyses to select isolates expressing thermo-regulated proteins of a predefined size. Selected isolates were then subjected to further phenotypic analyses such as Caco-2 cell adhesion and TEM. Isolates expressing putative CF proteins were subsequently chosen for genomic analyses using the already existing WGS data from Paper I. Four putative CF operons related to the AFA/Dr/AAF family of adhesins, which may be found in some EAEC, UPEC and DAEC strains, were identified; two of these operons were harbored by isolates part of L16 and two of them were harbored by a distantly related isolate. Mass spectrometric analysis of a thermo-regulated ~15 kD band, identified in the protein extract of one of the L16 isolates, revealed peptides matching the adhesin encoded by one of the AFA/Dr/AAF operons. Furthermore, another putative CF operon, which was highly similar to the operons encoding the porcine CF K88 and the human CF CS23, was identified. However, the putative major subunit was completely different from the major subunits in the operons encoding K88 and CS23. This finding indicates that this operon may encode a novel CF and suggests that the putative CF operon, K88 operon and CS23 operon, share a common ancestor and that the host-specificity most likely

is due to the sequence differences seen in the gene encoding the major subunits. The identification of putative CF operons, which may have been derived from different pathogenic *E. coli* and may be involved in virulence, supports that genetic exchange may lead to the formation of "hybrid" strains.

Further studies are planned to elucidate how a lineage, as those identified in Paper I, may be defined. For example, by using the whole genome sequences it is possible to identify the accessory genome of each major lineage using a clustering software (Roary) [122]. Such analysis will make it possible to analyze each lineage with even higher resolution, compared to only considering the core genes as was done in Paper I, and to identify lineage specific genes. In pathogenic enteropathogens, such as *Shigella spp.*, EIEC and EHEC, anti-virulence genes (AVGs), i.e. genes which are known to interfere with different virulence pathways and these genes subsequently have been lost over time or differentially expressed, have been identified [130,149,150]. Investigating ETEC lineages for the loss of such genes, which are present in ancestral isolates may aid in understanding why the major lineages have been stable over time.

When building the phylogenetic tree of the 362 ETEC isolates, recombination sites were removed since they may generate a false impression of the inter-strain relationships. Horizontal transfer of accessory genes (including virulence genes) is not the only event increasing the diversity of the *E. coli* species. Regulatory regions have also been shown to be transferred between related strains, and even across different bacterial species, resulting in major phenotypic changes or even the conversion of a commensal to a pathogen [151-153]. Exploring the role of horizontal regulatory switching may help us understand the diversity of ETEC even further. Two isolates each from the initial 11 identified lineages presented in Paper I have been PacBio sequenced. These sequences may be used in future studies as suitable reference genomes to identify the respective lineages, e.g. in genomic and epidemiological studies. The PacBio sequences may also be used to study the genomic content of each of the 11 lineages in more detail.

In addition to attempting to identify novel CFs, the WGS data presented in Paper I has been used in additional studies not part of these thesis, such as exploring the diversity of both the LT and ST toxins. Twelve new LT variants have been reported and shown to be associated with specific CF profiles

[154] and indirectly to the core genetic content. ST has been found to be less diverse than LT, mainly because of the size difference, but three new variants of STh and STp have been identified (Joffre *et al*. submitted manuscript). Furthermore, a recent study based on whole genome sequenced and PacBio sequenced ETEC isolates showed that the stability of CS6 is linked to the toxin profile and the presence of stability genes [155].

A new *in silico* methodology for identifying novel O genotypes in whole genome sequences of *E. coli* isolates has also been developed (Atsushi *et al*. in manuscript). In Paper I ten new O genotypes were identified using this method. The WGS data may also be used to elucidate the emergence and increase of antibiotic resistance genes over time, and such analyses are in progress. Using the large ETEC WGS database may also allow identification of conserved ETEC specific proteins, which may be components of a future ETEC vaccine.

The studies included in this thesis have been important for increasing the understanding of the evolution and population structure of ETEC and have re-defined the view of ETEC as a pathovar through the identification of lineages, revealing an association between the plasmid-encoded virulence genes and the chromosomal backbone. Furthermore, two promising approaches to identify putative novel ETEC CFs have been developed which resulted in the discovery of the novel CF CS30 and several putative CF operons. This continued search for novel virulence factors is important for our understanding of ETEC and increases our appreciation of its diversity and complexity.

ACKNOWLEDGEMENTS

My deepest thank you to everyone at the Department of Microbiology and Immunology at the University of Gothenburg and to the Infectious Genomics department at the Wellcome Trust Sanger Institute who've made this experience so much fun and for all the interesting people I have met.

Thousand thanks to:

Ann-Mari, my main supervisor – thanks to you I got the opportunity to take part in this exciting project. With your support, your positive spirit and your fantastic guidance nothing was impossible, at least that was how I felt. I have learnt so much from you and I look forward to keep doing that.

Åsa, my co-supervisor – thank you for almost always having time to discuss research, PCR, life goals, etc. When you moved to Stockholm I realized how much support you were for me, but Skype worked well too ☺

Gordon, my other co-supervisor – for believing in me and for taking good care of me whenever I visited. Thank you for being incredibly supportive and introducing me to so many fantastic researchers.

Gudrun – without you I would be lost in the lab! Thank you for lending a helping hand and support whenever I needed it, and for telling me when to slow down!

Nick – you have always been generous of your time, never saying no to discuss my projects and future plans. I appreciate your frankness and your inspirational persona. I look forward to continue working and learning from you.

Ingrid and **Matilda** – thanks for introducing me to ETEC and getting me hooked during my Bachelor thesis. Thank you Ingrid for always taking your time to discuss different matters.

Anna – I have always enjoyed our spontaneous discussions in the corridors and whenever you've given me advice, I've listened.

My roommates, **Samuel** and **Inta**, the best ones you could have. Inta, you have kept my sweet cravings at bay with supplying Latvian candies and Samuel, you have fed me with funny YouTube clips and rock! It's going to be amaze-balls to continue sharing office with you two loonies.

Susannah – for your help in proofreading this thesis, hopefully I have learnt to use commas in every sentence. Appreciate our spontaneous chats in the stairs, corridor and toilet, and for letting me blow of some steam when needed ☺

Stefan – for always having time to help me with this and that: inspiration for figures, to define definitions or not to, act as a sounding board and with all the help with word.

Paulina – for guiding me in the jungle of stipends and for all the "fikas"/lunches. I wish your calmness would rub of on me from time to time, but it didn't… You're up next!

Sean – för the help med the English language.

Lynda – for joining the ETEC-team. Maybe we could go and get pampered, for real, sometime!

Josh – for generating the CS30 mutant and for re-introducing me to EMiL =)

Frida – for your supportive words, they mean a lot, and for being you!

Susanne K – for all the interesting and fun chats we have. I am really happy you moved into the lab!

Mike and **Jan** – for helping me prepare for this day.

Ankur – thank you for making me feel at home when I arrived at Sanger, missed you so much after you left. I hope our roads will continue to cross in the future.

Del – for always having a smile on your face, which makes any problem feel solvable, and for sharing you knowledge on plasmids.

Jacqui, **Martin A**, **Andrew** and **Martin H** – my heroes! Without you I would be lost. You guided me in the new world of bioinformatics.

My childhood friend, **Johanna** – for being incredibly supportive. With such dedication and drive this ought to be the best party, couldn't have done it without you!

My dad, **Bengt** – for all the phone calls when you asked me if I had discovered something new or how the writing was going and for joining me to Washington DC and witnessed my first international scientific presentation.

My husband, **Claes** – for being there and supporting me despite my quirkiness.

My daughter, **Elle** – for always meeting me in the hallway with the biggest smile, you make every problem seem small and insignificant.

We gratefully acknowledge the following organizations for financial support of this thesis: The Swedish Research Council (grants 2012-3464 and 2011-3435), the Swedish Strategic Foundation (grant SB12-0072, the Wellcome Trust (grant 098051) and the Fru Mary von Sydows, född Wijk, Foundation.

REFERENCES

1. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clinical Microbiology Reviews. 2013;26: 822–880.

2. Moriel DG, Rosini R, Seib KL, Serino L, Pizza M, Rappuoli R. *Escherichia coli*: great diversity around a common core. MBio. 2012;3: 1-3.

3. Kaper JB, Nataro JP, Mobley HL. Pathogenic *Escherichia coli*. Nature Reviews Microbiology. 2004;2: 123–140.

4. Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, *et al.* Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. The Lancet. 2012;379: 2151–2161.

5. Qadri F, Svennerholm AM, Faruque AS, Sack RB. Enterotoxigenic *Escherichia coli* in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. Clinical Microbiology Reviews. 2005;18: 465–483.

6. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, *et al.* 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. The Lancet. 2013;382: 209–222.

7. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, *et al.* Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet. 2012;380: 2095–2128.

8. la Cabada Bauche de J, Dupont HL. New Developments in Traveler's Diarrhea. Gastroenterology & Hepatology. 2011;7: 88–95.

9. Steffen R, Hill DR, DuPont HL. Traveler's diarrhea: a clinical review. JAMA. 2015;313: 71-80.

10. Nada RA, Armstrong A, Shaheen HI, Nakhla I, Sanders JW, Riddle MS, *et al.* Phenotypic and genotypic characterization of enterotoxigenic *Escherichia coli* isolated from U.S. military personnel participating in Operation Bright Star, Egypt, from 2005 to 2009. Diagnostic Microbiology and Infectious Diseases. 2013;76: 272–277.

11. Nagy B, Fekete PZ. Enterotoxigenic *Escherichia coli* in veterinary medicine.

International Journal of Medical Microbiology. 2005;295: 443–454.

12. Svennerholm AM, Lundgren A. Recent progress toward an enterotoxigenic *Escherichia coli* vaccine. Expert Reviews. 2012;11: 495–507.

13. Porter CK, Riddle MS, Tribble DR, Louis Bougeois A, McKenzie R, Isidean SD, *et al.* A systematic review of experimental infections with enterotoxigenic *Escherichia coli* (ETEC). Vaccine. 2011;29: 5869–5885.

14. Gaastra W, Svennerholm AM. Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC). Trends in Microbiology. 1996;4: 444–452.

15. Clemens J, Shin S, Sur D, Nair GB, Holmgren J. New-generation vaccines against cholera. Nature Reviews: Gastroenterology & Hepatology. 2011;8: 701–710.

16. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. Clinical Microbiology Reviews. 1998;11: 142–201.

17. Rao MC. Toxins which activate guanylate cyclase: heat-stable enterotoxins. Ciba Foundation Symposium 112 - Microbial Toxins and Diarrhoeal Disease. 1985. pp. 74–93.

18. Tien X-Y, Brasitus TA, Kaetzel MA, Dedman JR, Nelson DJ. Activation of the cystic fibrosis transmembrane conductance regulator by cGMP in the human colonic cancer cell line, Caco-2. The Journal of Biological Chemistry. 1994;269: 51–54.

19. Holmgren J, Svennerholm A-M. Immunological cross-reactivity between *Escherichia coli* heat-labile enterotoxin and cholera toxin A and B subunits. Current Microbiology. 1979;2: 55–58.

20. Rao MR, Abu Elyazeed R, Savarino SJ, Naficy AB, Wierzba TF, Abdel-Messih I, *et al.* High disease burden of diarrhea due to enterotoxigenic *Escherichia coli* among rural Egyptian infants and young children. Journal of Clinical Microbiology. 2003;41: 4862–4864.

21. Handl CE, Olsson E, Flock JI. Evaluation of three different STb assays and comparison of enterotoxin pattern over a five-year period in Swedish porcine *Escherichia coli*. Diagnostic Microbiology and Infectious Diseases. 1992;15: 505–510.

22. Bolin I, Wiklund G, Qadri F, Torres O, Bourgeois AL, Savarino S, *et al.* Enterotoxigenic *Escherichia coli* with STh and STp genotypes is associated with diarrhea both in children in areas of endemicity and in travelers. Journal

of Clinical Microbiology. 2006;44: 3872–3877.

23. Tauschek M, Gorrell RJ, Strugnell RA, Robins-Browne RM. Identification of a protein secretory pathway for the secretion of heat-labile enterotoxin by an enterotoxigenic strain of *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America. 2002;99: 7066–7071.

24. Yamanaka H, Nomura T, Fujii Y, Okamoto K. Need for TolC, an *Escherichia coli* outer membrane protein, in the secretion of heat-stable enterotoxin I across the outer membrane. Microbial Pathogenesis. 1998;25: 111–120.

25. Wensink J, Gankema H, Jansen WH, Guinée PA, Witholt B. Isolation of the membranes of an enterotoxigenic strain of *Escherichia coli* and distribution of enterotoxin activity in different subcellular fractions. Biochimica et Biophysica Acta. 1978;514: 128–136.

26. Gyles C, So M, Falkow S. The enterotoxin plasmids of *Escherichia coli*. The Journal of Infectious Diseases. 1974;130: 40–49.

27. Wolf MK. Occurrence, distribution, and associations of O and H serogroups, colonization factor antigens, and toxins of enterotoxigenic *Escherichia coli*. Clinical Microbiology Reviews. 1997;10: 569–584.

28. Nuccio SP, Baumler AJ. Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. Microbiology and Molecular Biology Reviews. 2007;71: 551–575.

29. Wurpel DJ, Beatson SA, Totsika M, Petty NK, Schembri MA. Chaperone-usher fimbriae of *Escherichia coli*. PLoS One. 2013;8: e52835.

30. Hirst TR, Sanchez J, Kaper JB, Hardy SJ, Holmgren J. Mechanism of toxin secretion by *Vibrio cholerae* investigated in strains harboring plasmids that encode heat-labile enterotoxins of *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America. 1984;81: 7752–7756.

31. Wandersman C. Secretion across the bacterial outer membrane. Trends in Genetics. 1992;8: 317–322.

32. Kudva R, Denks K, Kuhn P, Vogt A, Müller M, Koch H-G. Protein translocation across the inner membrane of Gram-negative bacteria: the Sec and Tat dependent protein transport pathways. Research in Microbiology. 2013;164: 505–534.

33. Ng TW, Akman L, Osisami M, Thanassi DG. The usher N terminus is the initial targeting site for chaperone-subunit complexes and participates in

subsequent pilus biogenesis events. Journal of Bacteriology. 2004;186: 5321–5331.

34. Thanassi DG. Ushers and secretins: channels for the secretion of folded proteins across the bacterial outer membrane. Journal of Molecular Microbiology and Biotechnology. 2002;4: 11–20.

35. Isidean SD, Riddle MS, Savarino SJ, Porter CK. A systematic review of ETEC epidemiology focusing on colonization factor and toxin expression. Vaccine. 2011;29: 6167–6178.

36. Tobias J, Svennerholm A-M. Strategies to overexpress enterotoxigenic *Escherichia coli* (ETEC) colonization factors for the construction of oral whole-cell inactivated ETEC vaccine candidates. Applied Microbiology and Biotechnology. 2012;93: 2291–2300. 3930-6

37. Anantha RP, McVeigh AL, Lee LH, Agnew MK, Cassels FJ, Scott DA, *et al.* Evolutionary and functional relationships of colonization factor antigen I and other class 5 adhesive fimbriae of enterotoxigenic *Escherichia coli*. Infection and Immunity. 2004;72: 7190–7201.

38. Nada RA, Shaheen HI, Khalil SB, Mansour A, El-Sayed N, Touni I, *et al.* Discovery and phylogenetic analysis of novel members of class b enterotoxigenic *Escherichia coli* adhesive fimbriae. Journal of Clinical Microbiology. 2011;49: 1403–1410.

39. Evans DG, Evans DJ, Tjoa WS, DuPont HL. Detection and characterization of colonization factor of enterotoxigenic *Escherichia coli* isolated from adults with diarrhea. Infection and Immunity. 1978;19: 727–736.

40. Jordi BJ, Willshaw GA, van der Zeijst BA, Gaastra W. The complete nucleotide sequence of region 1 of the CFA/I fimbrial operon of human enterotoxigenic *Escherichia coli*. DNA Sequence: The Jorunal of DNA Sequencing and Mapping. 1992;2: 257–263.

41. Evans DG, Evans DJ. New surface-associated heat-labile colonization factor antigen (CFA/II) produced by enterotoxigenic *Escherichia coli* of serogroups O6 and O8. Infection and Immunity. 1978;21: 638–647.

42. Marron MB, Smyth CJ. Molecular analysis of the *cso* operon of enterotoxigenic *Escherichia coli* reveals that CsoA is the adhesin of CS1 fimbriae and that the accessory genes are interchangeable with those of the *cfa* operon. Microbiology. 1995;141: 2849–2859.

43. Smyth CJ. Two mannose-resistant haemagglutinins on enterotoxigenic

Escherichia coli of serotype O6:K15:H16 or H-isolated from travellers' and infantile diarrhoea. Journal of General Microbiology. 1982;128: 2081–2096.

44. Froehlich BJ, Karakashian A, Sakellaris H, Scott JR. Genes for CS2 pili of enterotoxigenic *Escherichia coli* and their interchangeability with those for CS1 pili. Infection and Immunity. 1995;63: 4849–4856.

45. Thomas LV, McConnell MM, Rowe B, Field AM. The possession of three novel coli surface antigens by enterotoxigenic *Escherichia coli* strains positive for the putative colonization factor PCF8775. Journal of General Microbiology. 1985;131: 2319–2326.

46. McConnell MM, Chart H, Field AM, Hibberd M, Rowe B. Characterization of a putative colonization factor (PCFO166) of enterotoxigenic *Escherichia coli* of serogroup O166. Journal of General Microbiology. 1989;135: 1135–1144.

47. McConnell MM, Hibberd M, Field AM, Chart H, Rowe B. Characterization of a new putative colonization factor (CS17) from a human enterotoxigenic *Escherichia coli* of serotype O114:H21 which produces only heat-labile enterotoxin. The Journal of Infectious Diseases. 1990;161: 343–347.

48. Chattopadhyay S, Tchesnokova V, McVeigh A, Kisiela DI, Dori K, Navarro A, *et al.* Adaptive evolution of class 5 fimbrial genes in enterotoxigenic *Escherichia coli* and its functional consequences. The Journal of Biological Chemistry. 2012;287: 6150–6158.

49. Clark CA, Heuzenroeder MW, Manning PA. Colonization factor antigen CFA/IV (PCF8775) of human enterotoxigenic *Escherichia coli*: nucleotide sequence of the CS5 determinant. Infection and Immunity. 1992;60: 1254–1257.

50. Hibberd ML, McConnell MM, Field AM, Rowe B. The fimbriae of human enterotoxigenic *Escherichia coli* strain 334 are related to CS5 fimbriae. Journal of General Microbiology. 1990;136: 2449–2456.

51. Tacket CO, Maneval DR, Levine MM. Purification, morphology, and genetics of a new fimbrial putative colonization factor of enterotoxigenic *Escherichia coli* O159:H4. Infection and Immunity. 1987;55: 1063–1069.

52. Viboud GI, Binsztein N, Svennerholm AM. A new fimbrial putative colonization factor, PCFO20, in human enterotoxigenic *Escherichia coli*. Infection and Immunity. 1993;61: 5190–5197.

53. Valvatne H, Sommerfelt H, Gaastra W, Bhan MK, Grewal HM.

Identification and characterization of CS20, a new putative colonization factor of enterotoxigenic *Escherichia coli*. Infection and Immunity. 1996;64: 2635–2642.

54. Jalajakumari MB, Thomas CJ, Halter R, Manning PA. Genes for biosynthesis and assembly of CS3 pili of CFA/II enterotoxigenic *Escherichia coli*: novel regulation of pilus production by bypassing an amber codon. Molecular Microbiology. 1989;3: 1685–1695.

55. McConnell MM, Thomas LV, Scotland SM, Rowe B. The possession of coli surface antigen CS6 by enterotoxigenic *Escherichia coli* of serogroups O25, O27, O148, and O159: a possible colonization factor? Current Microbiology. 1986;14: 51–54.

56. Svennerholm AM, Vidal YL, Holmgren J, McConnell MM, Rowe B. Role of PCF8775 antigen and its coli surface subcomponents for colonization, disease, and protective immunogenicity of enterotoxigenic *Escherichia coli* in rabbits. Infection and Immunity. 1988;56: 523–528.

57. Taniguchi T, Fujino Y, Yamamoto K, Miwatani T, Honda T. Sequencing of the gene encoding the major pilin of pilus colonization factor antigen III (CFA/III) of human enterotoxigenic *Escherichia coli* and evidence that CFA/III is related to type IV pili. Infection and Immunity. 1995;63: 724–728.

58. Giron JA, Levine MM, Kaper JB. Longus: a long pilus ultrastructure produced by human enterotoxigenic *Escherichia coli*. Molecular Microbiology. 1994;12: 71–82.

59. Aubel D, Darfeuille-Michaud A, Joly B. New adhesive factor (antigen 8786) on a human enterotoxigenic *Escherichia coli* O117:H4 strain isolated in Africa. Infection and Immunity. 1991;59: 1290–1299.

60. Pichel M, Binsztein N, Viboud G. CS22, a novel human enterotoxigenic *Escherichia coli* adhesin, is related to CS15. Infection and Immunity. 2000;68: 3280–3285.

61. Forestier C, Welinder KG, Darfeuille-Michaud A, Klemm P. Afimbrial adhesin from *Escherichia coli* strain 2230: Purification, characterization and partial covalent structure. FEMS Microbiology Letters. 1987;40: 47–50.

62. Knutton S, Lloyd DR, McNeish AS. Identification of a new fimbrial structure in enterotoxigenic *Escherichia coli* (ETEC) serotype O148:H28 which adheres to human intestinal mucosa: a potentially new human ETEC colonization factor. Infection and Immunity. 1987;55: 86–92.

63. Heuzenroeder MW, Elliot TR, Thomas CJ, Halter R, Manning PA. A new fimbrial type (PCFO9) on enterotoxigenic *Escherichia coli* 09:H- LT+ isolated from a case of infant diarrhea in central Australia. FEMS Microbiology Letters. 1990;54: 55–60.

64. Del Canto F, Botkin DJ, Valenzuela P, Popov V, Ruiz-Perez F, Nataro JP, *et al.* Identification of coli surface antigen 23, a novel adhesin of enterotoxigenic *Escherichia coli*. Infection and Immunity. 2012;80: 2791–2801.

65. Caron J, Coffield LM, Scott JR. A plasmid-encoded regulatory gene, *rns*, required for expression of the CS1 and CS2 adhesins of enterotoxigenic *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America. 1989;86: 963–967.

66. Wennerås C, Holmgren J, Svennerholm AM. The binding of colonization factor antigens of enterotoxigenic *Escherichia coli* to intestinal cell membrane proteins. FEMS Microbiology Letters. 1990;54: 107–112.

67. Pieroni P, Worobec EA, Paranchych W, Armstrong GD. Identification of a human erythrocyte receptor for colonization factor antigen I pili expressed by H10407 enterotoxigenic *Escherichia coli*. Infection and Immunity. 1988;56: 1334–1340.

68. Jansson L, Tobias J, Lebens M, Svennerholm A-M, Teneberg S. The major subunit, CfaB, of colonization factor antigen I from enterotoxigenic *Escherichia coli* is a glycosphingolipid binding protein. Infection and Immunity. 2006;74: 3488–3497.

69. Ahmed T, Lundgren A, Arifuzzaman M, Qadri F, Teneberg S, Svennerholm A-M. Children with the Le(a+b-) blood group have increased susceptibility to diarrhea caused by enterotoxigenic *Escherichia coli* expressing colonization factor I group fimbriae. Infection and Immunity. 2009;77: 2059–2064.

70. Orø HS, Kolstø AB, Wennerås C, Svennerholm AM. Identification of asialo-GM1 as a binding structure for *Escherichia coli* colonization factor antigens. FEMS Microbiology Letters. 1990;60: 289–292.

71. Jansson L, Tobias J, Jarefjäll C, Lebens M, Svennerholm A-M, Teneberg S. Sulfatide recognition by colonization factor antigen CS6 from enterotoxigenic *Escherichia coli*. PLoS One. 2009;4: e4487.

72. Harris AM, Chowdhury F, Begum YA, Khan AI, Faruque ASG, Svennerholm A-M, *et al.* Shifting prevalence of major diarrheal pathogens in patients seeking hospital care during floods in 1998, 2004, and 2007 in Dhaka, Bangladesh. The American Society of Tropical Medicine and Hygiene.

2008;79: 708–714.

73. Shaheen HI, Abdel Messih IA, Klena JD, Mansour A, El-Wakkeel Z, Wierzba TF, *et al.* Phenotypic and genotypic analysis of enterotoxigenic *Escherichia coli* in samples obtained from Egyptian children presenting to referral hospitals. Journal of Clinical Microbiology. 2009;47: 189–197.

74. Gonzales L, Sanchez S, Zambrana S, Iñiguez V, Wiklund G, Svennerholm A-M, *et al.* Molecular characterization of enterotoxigenic *Escherichia coli* isolates recovered from children with diarrhea during a 4-year period (2007 to 2010) in Bolivia. Journal of Clinical Microbiology. 2013;51: 1219–1225.

75. Begum YA, Baby NI, Faruque ASG, Jahan N, Cravioto A, Svennerholm A-M, *et al.* Shift in phenotypic characteristics of enterotoxigenic *Escherichia coli* (ETEC) isolated from diarrheal patients in Bangladesh. PLoS Neglected Tropical Diseases. 2014;8: e3031.

76. Gomez-Duarte OG, Chattopadhyay S, Weissman SJ, Giron JA, Kaper JB, Sokurenko EV. Genetic diversity of the gene cluster encoding longus, a type IV pilus of enterotoxigenic *Escherichia coli*. Journal of Bacteriology. 2007;189: 9145–9149.

77. Njoroge SM, Boinett CJ, Madé LF, Ouko TT, Fèvre EM, Thomson NR, *et al.* A putative, novel coli surface antigen 8B (CS8B) of enterotoxigenic *Escherichia coli*. FEMS Pathogens and Disease. 2015;73: ftv047.

78. Svennerholm AM, Tobias J. Vaccines against enterotoxigenic *Escherichia coli*. Expert Review of Vaccines. 2008;7: 795–804.

79. Nataro JP, Deng Y, Maneval DR, German AL, Martin WC, Levine MM. Aggregative adherence fimbriae I of enteroaggregative *Escherichia coli* mediate adherence to HEp-2 cells and hemagglutination of human erythrocytes. Infection and Immunity. 1992;60: 2297–2304.

80. Madhavan TPV, Sakellaris H. Colonization factors of enterotoxigenic *Escherichia coli*. Advanced Applied Microbiology. 2015;90: 155–197.

81. Nowicki B, Selvarangan R, Nowicki S. Family of *Escherichia coli* Dr adhesins: decay-accelerating factor receptor recognition and invasiveness. The Journal of Infectious Diseases. 2001;183 Suppl 1: S24–7.

82. Servin AL. Pathogenesis of human diffusely adhering *Escherichia coli* expressing Afa/Dr adhesins (Afa/Dr DAEC): current insights and future challenges. Clinical Microbiology Reviews. 2014;27: 823–869.

83. Jønsson R, Struve C, Boisen N, Mateiu RV, Santiago AE, Jenssen H, *et al.* Novel aggregative adherence fimbria variant of enteroaggregative *Escherichia coli*. Infection and Immunity. 2015;83: 1396–1405.

84. Soto GE, Hultgren SJ. Bacterial adhesins: common themes and variations in architecture and assembly. Journal of Bacteriology. 1999;181: 1059–1071.

85. Wagner C, Hensel M. Adhesive mechanisms of *Salmonella enterica*. Advances in Experimental Medicine and Biology. 2011;715: 17–34.

86. Farfan MJ, Cantero L, Vidal R, Botkin DJ, Torres AG. Long polar fimbriae of enterohemorrhagic *Escherichia coli* O157:H7 bind to extracellular matrix proteins. Infection and Immunity. 2011;79: 3744–3750.

87. Fleckenstein J, Sheikh A, Qadri F. Novel antigens for enterotoxigenic *Escherichia coli* vaccines. Expert Review of Vaccines. 2014;13: 631–639.

88. Luo Q, Kumar P, Vickers TJ, Sheikh A, Lewis WG, Rasko DA, *et al.* Enterotoxigenic *Escherichia coli* secretes a highly conserved mucin-degrading metalloprotease to effectively engage intestinal epithelial cells. Infection and Immunity. 2014;82: 509–521.

89. Chen Q, Savarino SJ, Venkatesan MM. Subtractive hybridization and optical mapping of the enterotoxigenic *Escherichia coli* H10407 chromosome: isolation of unique sequences and demonstration of significant similarity to the chromosome of *E. coli* K-12. Microbiology. 2006;152: 1041–1054.

90. Ouyang Z, Isaacson R. Identification and characterization of a novel ABC iron transport system, *fit*, in *Escherichia coli*. Infection and Immunity. 2006;74: 6949–6956.

91. Nesta B, Spraggon G, Alteri C, Moriel DG, Rosini R, Veggi D, *et al.* FdeC, a novel broadly conserved *Escherichia coli* adhesin eliciting protection against urinary tract infections. MBio. 2012;3: e00010–12.

92. Patel SK, Dotson J, Allen KP, Fleckenstein JM. Identification and molecular characterization of EatA, an autotransporter protein of enterotoxigenic *Escherichia coli*. Infection and Immunity. 2004;72: 1786–1794.

93. Sjöling A, von Mentzer A, Svennerholm A-M. Implications of enterotoxigenic *Escherichia coli* genomics for vaccine development. Expert Review of Vaccines. 2015;14: 551–560.

94. Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, *et al.* A comparative genomic analysis of diverse clonal types of enterotoxigenic

Escherichia coli reveals pathovar-specific conservation. Infection and Immunity. 2011;79: 950–960.

95. Pilonieta MC, Bodero MD, Munson GP. CfaD-dependent expression of a novel extracytoplasmic protein from enterotoxigenic *Escherichia coli*. Journal of Bacteriology. 2007;189: 5060–5067.

96. Elsinghorst EA, Weitz JA. Epithelial cell invasion and adherence directed by the enterotoxigenic *Escherichia coli tib* locus is associated with a 104-kilodalton outer membrane protein. Infection and Immunity. 1994;62: 3463–3471.

97. Fleckenstein JM, Kopecko DJ, Warren RL, Elsinghorst EA. Molecular characterization of the *tia* invasion locus from enterotoxigenic *Escherichia coli*. Infection and Immunity. 1996;64: 2256–2265.

98. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 1977;74: 5463–5467.

99. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, *et al.* The complete genome sequence of *Escherichia coli* K-12. Science. 1997;277: 1453–1462.

100. Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. Nature Reviews Microbiology. 2015;13:787-794.

101. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010;327: 469–474.

102. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, *et al.* Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. Nature Genetics. 2012;44: 1215–1221.

103. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. Nature Genetics. 2013;45: 656–663.

104. Bryant JM, Schurch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, *et al.* Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infectious Diseases. 2013;13: 110.

105. Holt KE, Baker S, Weill FX, Holmes EC, Kitchen A, Yu J, *et al. Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. Nature Genetics. 2012;44: 1056–1059.

106. Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011;477: 462–465.

107. Moriel DG, Bertoldi I, Spagnuolo A, Marchi S, Rosini R, Nesta B, *et al.* Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. Proceedings of the National Academy of Sciences of the United States of America. 2010;107: 9072–9077.

108. Roy K, Bartels S, Qadri F, Fleckenstein JM. Enterotoxigenic *Escherichia coli* elicits immune responses to multiple surface proteins. Infection and Immunity. 2010;78: 3027–3035.

109. Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M, Cunningham AF, *et al.* A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407. Journal of Bacteriology. 2010;192: 5822–5831.

110. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. Journal of Bacteriology. 2008;190: 6881–6893.

111. Steinsland H, Lacher DW, Sommerfelt H, Whittam TS. Ancestral lineages of human enterotoxigenic *Escherichia coli*. Journal of Clinical Microbiology. 2010;48: 2916–2924.

112. Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur E. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. Molecular Biology and Evolution. 2004;21: 1085–1094.

113. Turner SM, Chaudhuri RR, Jiang ZD, DuPont H, Gyles C, Penn CW, *et al.* Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. Journal of Clinical Microbiology. 2006;44: 4528–4536.

114. Sjöling A, Wiklund G, Savarino SJ, Cohen DI, Svennerholm AM. Comparative analyses of phenotypic and genotypic methods for detection of enterotoxigenic *Escherichia coli* toxins and colonization factors. Journal of

Clinical Microbiology. 2007;45: 3295–3301.

115. Rodas C, Iniguez V, Qadri F, Wiklund G, Svennerholm AM, Sjöling A. Development of multiplex PCR assays for detection of enterotoxigenic *Escherichia coli* colonization factors and toxins. Journal of Clinical Microbiology. 2009;47: 1218–1220.

116. Nicklasson M, Sjöling A, von Mentzer A, Qadri F, Svennerholm AM. Expression of colonization factor CS5 of enterotoxigenic *Escherichia coli* (ETEC) is enhanced in vivo and by the bile component Na glycocholate hydrate. PLoS One. 2012;7: e35827.

117. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008;18: 821–829.

118. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. The ISME Journal. 2009;3: 199–208.

119. Martincorena I, Seshasayee AS, Luscombe NM. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature. 2012;485: 95–98.

120. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Molecular Biology and Evolution. 2013;30: 1224–1228.

121. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research. 2011;39: W29–37.

122. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, *et al.* Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015.

123. Viboud GI, McConnell MM, Helander A, Svennerholm AM. Binding of enterotoxigenic *Escherichia coli* expressing different colonization factors to tissue-cultured Caco-2 cells and to isolated human enterocytes. Microbial Pathogenesis. 1996;21: 139–147.

124. Del Canto F, Valenzuela P, Cantero L, Bronstein J, Blanco JE, Blanco J, *et al.* 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. Journal of Clinical Microbiology. 2011;49: 3198–3203.

125. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, *et al.*

Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences of the United States of America. 1998;95: 3140–3145.

126. Duriez P, Clermont O, Picard B, Denamur E, Bonacorsi S, Bingen E, *et al.* Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. Microbiology. 2001;147: 1671–1676.

127. Sahl JW, Sistrunk JR, Fraser CM, Hine E, Baby N, Begum Y, *et al.* Examination of the enterotoxigenic *Escherichia coli* population structure during human infection. MBio. 2015;6: e00501.

128. Njamkepo E, Fawal N, Tran-Dien A, Hawkey J, Strockbine N, Jenkins C, *et al.* Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1. Nature Microbiology. 2016;16027.

129. Connor TR, Barker CR, Baker KS, Weill F-X, Talukder KA, Smith AM, *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. Elife. 2015;4: e07335.

130. Bliven KA, Maurelli AT. Antivirulence genes: insights into pathogen evolution through gene loss. Infection and Immunity. 2012;80: 4061–4070.

131. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, *et al.* Rapid pneumococcal evolution in response to clinical interventions. Science. 2011;331: 430–434.

132. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, *et al.* Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. Genome Biology. 2012;13: R118.

133. Rivera FP, Ochoa TJ, Maves RC, Bernal M, Medina AM, Meza R, *et al.* Genotypic and phenotypic characterization of enterotoxigenic *Escherichia coli* strains isolated from Peruvian children. Journal of Clinical Microbiology. 2010;48: 3198–3203.

134. Honarvar S, Choi B-K, Schifferli DM. Phase variation of the 987P-like CS18 fimbriae of human enterotoxigenic *Escherichia coli* is regulated by site-specific recombinases. Molecular Microbiology. 2003;48: 157–171.

135. Gonzales L, Ali ZB, Nygren E, Wang Z, Karlsson S, Zhu B, *et al.* Alkaline pH Is a signal for optimal production and secretion of the heat labile toxin, LT in enterotoxigenic *Escherichia coli* (ETEC). PLoS One. 2013;8: e74069.

136. Haycocks JRJ, Sharma P, Stringer AM, Wade JT, Grainger DC. The molecular basis for control of ETEC enterotoxin expression in response to environment and host. PLoS Pathogens. 2015;11: e1004605.

137. McCarthy A. Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. Chemistry & Biology. 2010;17: 675–676.

138. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biology. 2015;16: 294.

139. Flett F, Humphreys GO, Saunders JR. Intraspecific and intergeneric mobilization of non-conjugative resistance plasmids by a 24.5 megadalton conjugative plasmid of *Neisseria gonorrhoeae*. Journal of General Microbiology. 1981;125: 123–129.

140. Udo EE, Love H, Grubb WB. Intra- and inter-species mobilisation of non-conjugative plasmids in staphylococci. Journal of Medical Microbiology. 1992;37: 180–186.

141. Meyer R. Replication and conjugative mobilization of broad host-range IncQ plasmids. Plasmid. 2009;62: 57–70.

142. O'Brien FG, Yui Eto K, Murphy RJT, Fairhurst HM, Coombs GW, Grubb WB, *et al.* Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. Nucleic Acids Research. 2015;43: 7971–7983.

143. Boisen N, Struve C, Scheutz F, Krogfelt KA, Nataro JP. New adhesin of enteroaggregative *Escherichia coli* related to the Afa/Dr/AAF family. Infection and Immunity. 2008;76: 3281–3292.

144. Bernier C, Gounon P, Le Bouguénec C. Identification of an aggregative adhesion fimbria (AAF) type III-encoding operon in enteroaggregative *Escherichia coli* as a sensitive probe for detecting the AAF-encoding operon family. Infection and Immunity. 2002;70: 4302–4311.

145. Bien J, Sokolova O, Bozko P. Role of uropathogenic *Escherichia coli* virulence factors in development of urinary tract infection and kidney damage. International Journal of Nephrology. 2012;2012: 681473–15.

146. Guignot J, Peiffer I, Bernet-Camard MF, Lublin DM, Carnoy C, Moseley SL, *et al.* Recruitment of CD55 and CD66e brush border-associated glycosylphosphatidylinositol-anchored proteins by members of the Afa/Dr diffusely adhering family of *Escherichia coli* that infect the human polarized

intestinal Caco-2/TC7 cells. Infection and Immunity. 2000;68: 3554–3563.

147. Heine M, Nollau P, Masslo C, Nielsen P, Freund B, Bruns OT, *et al.* Investigations on the usefulness of CEACAMs as potential imaging targets for molecular imaging purposes. PLoS One. 2011;6: e28030.

148. Huisman TT, Bakker D, Klaasen P, de Graaf FK. Leucine-responsive regulatory protein, IS1 insertions, and the negative regulator FaeA control the expression of the *fae* (K88) operon in *Escherichia coli*. Molecular Microbiology. 1994;11: 525–536.

149. Day WA, Fernández RE, Maurelli AT. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella spp.* Infection and Immunity. 2001;69: 7471–7480.

150. Vazquez-Juarez RC, Kuriakose JA, Rasko DA, Ritchie JM, Kendall MM, Slater TM, *et al.* CadA negatively regulates *Escherichia coli* O157:H7 adherence and intestinal colonization. Infection and Immunity. 2008;76: 5072–5081.

151. Lee DH, Palsson BØ. Adaptive evolution of *Escherichia coli* K-12 MG1655 during growth on a non-native carbon source, L-1,2-propanediol. Applied and Environmental Microbiology. 2010;76: 6327–6327.

152. Somvanshi VS, Sloup RE, Crawford JM, Martin AR, Heidt AJ, Kim K-S, *et al.* A single promoter inversion switches *Photorhabdus* between pathogenic and mutualistic states. Science. 2012;337: 88–93.

153. Oren Y, Smith MB, Johns NI, Kaplan Zeevi M, Biran D, Ron EZ, *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. Proceedings of the National Academy of Sciences of the United States of America. 2014;111: 16112–16117.

154. Joffre E, von Mentzer A, Abd El Ghany M, Oezguen N, Savidge T, Dougan G, *et al.* Allele variants of enterotoxigenic *Escherichia coli* heat-labile toxin are globally transmitted and associated with colonization factors. Journal of Bacteriology. 2015;197: 392–403.