Long-Read-Sequenced Reference Genomes of the Seven Major Lineages of Enterotoxigenic Escherichia Coli (ETEC) Circulating in Modern Time

Astrid von Mentzer (avm@sanger.ac.uk)
Wellcome Sanger Institute
Grace A Blackwell
European Bioinformatics Institute
Derek Pickard
University of Cambridge
Christine J Boinett
Wellcome Sanger Institute
Enrique Joffré
Karolinska Institute
Andrew J Page
Quadram Institute
Ann-Mari Svennerholm
University of Gothenburg
Gordon Dougan
University of Cambridge
Åsa Sjöling
Karolinska Institute

Research Article

Keywords: Reference genome, plasmid, enterotoxigenic Escherichia coli, ETEC, diarrhoea

Posted Date: February 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-237525/v1

Version of Record: A version of this preprint was published at Scientific Reports on April 29th, 2021. See the published version at https://doi.org/10.1038/s41598-021-88316-2.

Long-read-sequenced reference genomes of the seven major lineages of enterotoxigenic *Escherichia coli* (ETEC) circulating in modern time

Astrid von Mentzer[1,2*#], Grace A. Blackwell[1,3], Derek Pickard[4], Christine J. Boinett[1], Enrique Joffré[5], Andrew J Page[1,6], Ann-Mari Svennerholm[2], Gordon Dougan[4] and Åsa Sjöling[5]

[1]Wellcome Sanger Institute, Hinxton, Cambridge, UK

[2] Department of Microbiology and Immunology, Sahlgrenska Academy, University of Gothenburg, Sweden

[3] EMBL-EBI, Hinxton, Cambridge, UK

[4] University of Cambridge, Department of Medicine, Cambridge, UK

[5] Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet

[6] Quadram Institute Bioscience, Norwich Research Park, Norwich, UK

*corresponding author: avm@sanger.ac.uk or mentzerv@chalmers.se

[#]Currently affiliated with Chalmers University of Technology, Gothenburg, Sweden

Abstract

Enterotoxigenic *Escherichia coli* (ETEC) is an enteric pathogen responsible for the majority of diarrheal cases worldwide. ETEC infections are estimated to cause 80,000 fatalities per year, with the highest rates of burden, ca 75 million cases per year, amongst children under five years of age in resource-poor countries. It is also the leading cause of diarrhoea in travellers. Previous large-scale sequencing studies have found seven major ETEC lineages currently in circulation worldwide.

We used PacBio long-read sequencing combined with Illumina sequencing to create high-quality complete reference genomes for each of the major lineages with manually curated chromosomes and plasmids. We confirm that the major ETEC lineages all harbour conserved plasmids that have been associated with their respective background genomes for decades and that the plasmids and chromosomes of ETEC are both crucial for ETEC virulence and success as pathogens. The in-depth analysis of gene content, synteny and correct annotations of plasmids will elucidate other plasmids with and without virulence factors in related bacterial species. These reference genomes allow for fast and accurate comparison between different ETEC strains, and these data will form the foundation of ETEC genomics research for years to come.

Keywords: Reference genome, plasmid, enterotoxigenic *Escherichia coli*, ETEC, diarrhoea

Introduction

Diarrheal pathogens are a leading cause of morbidity and mortality globally (WHO, 2017), with enterotoxigenic *Escherichia coli* (ETEC) accounting for a large proportion of the diarrhoea cases in resource-poor countries [1]. An estimation of 220 million cases each year are attributed to ETEC (WHO PPC 2020). The most vulnerable group is children under five years, but ETEC can also cause disease in adults and is the principal cause of diarrhoea in travellers. Resource-poor settings, where access to clean water is limited, enable the spread of ETEC, transmitted via the faecal-oral route through ingestion of contaminated food or water [2]. The disease severity may range from mild to cholera-like symptoms with profuse watery diarrhoea. The infection is usually self-limiting, lasting three to four days and may be treated by water and electrolyte

rehydration to balance the loss of fluids and ions. There is strong evidence to support that an ETEC vaccine is of key importance to prevent children and adults from developing ETEC disease [3]. Several efforts are on-going to develop an ETEC vaccine, with the majority focusing on including immunogenic antigens possibly capable of inducing protection against a majority of the circulating ETEC clones [4–6].

ETEC bacteria adhere to the small intestine through fimbrial, fibrillar or afimbrial outer membrane-structures called colonisation factors (CF). Upon colonisation, the bacteria proliferate and secrete heat-labile toxin (LT) and/or heat-stable toxins, (STh or STp) causing diarrhoea and often vomiting causing the further spread of the bacteria in the environment [7].

The ability of an ETEC strain to infect relies on its ability to adhere to cells of a specific host. To this day, 27 different CFs with human tropism have been described, and individual ETEC strains usually express 1-3 different CFs [8–14]. The enterotoxins, LT and ST, can also be subdivided based on structure and function. Human-associated ETEC strains express one of the 28 different LT-I variants (LTh-1 and LTh-2 are the most common variants) [15] alone or together with one of the genetically distinct types of STa; STh and STp [16,17].

We have previously shown that ETEC strains causing human disease can be grouped into a set of clonal lineages that encompass strains with specific virulence profiles. Seven of the 21 identified lineages encompass ETEC strains that express the most commonly found CFs and toxin profiles amongst isolated clinical ETEC strains [4,18].

There is currently one complete ETEC reference genome, H10407 [19], with curated annotations. Several additional complete ETEC genomes are available [20,21], some of which are annotated using automated annotation pipelines that often fail at correctly annotating ETEC specific genes such as CFs. The rapid adaptation of next-generation sequencing in public health, specifically within bacterial diseases [22,23] and several large-scale sequencing studies [18,24–28] has led to a sharp increase in the number of publicly available ETEC genomes. Most of these data were generated with short-read technologies, such as Illumina. A limitation of short-read sequence data is the inability to unambiguously resolve repetitive regions of a genome, leading to fragmented de novo assemblies of the underlying genome, missing regions and genes, and disjointed synteny. ETEC is a highly diverse pathogen both in the core genome (SNPs) and the accessory genome, including mobile genetic elements (MGE). Clinically related MGEs, such as virulence plasmids, vary within ETEC strains. Hence, it is important to identify lineage-specific reference genomes that are carefully annotated, i.e. manually curated annotations, and include both chromosome and plasmid(s). Several complete genomes have been generated using long-read sequencing alone [20,28], however, circularising some chromosomes and plasmids may be difficult, and small plasmids can be lost. Assembly issues can be resolved using a hybrid assembly approach combining long-read and short-read sequencing data. In this report, we describe eight genomes, eight chromosomes (seven successfully circularised) and 29 plasmids (24 successfully circularised) with curated annotations, from isolates representing the major ETEC lineages (L1-L7) that cause disease globally. They are sequenced using both short and long-read sequencing technologies to provide the highest accuracy currently available. These reference genomes will form the foundation of ETEC genomics research for years to come.

Results

Genome analysis of eight representative ETEC isolates

Eight ETEC strains representing the seven major ETEC lineages (L1-L7) comprising isolates with the most prevalent virulence factor profiles were sequenced, assembled, circularised and manually curated (Table 1).

L3 includes two different representative strains, one CS7 and one CFA/I positive strain. All chromosomes except one were circularised (E1779). The average length of the chromosome was 4,927,521 bases (4,721,269-5,151,162) with an average GC content of 50.7% (50.4% to 50.9%) and the number of CDS ranging from 4,409 to 4,924 (Table S1). Each ETEC reference genome contains between two and five plasmids encompassing plasmid-specific features. Some of which carried virulence genes and/or antibiotic resistance genes (Table 2, Additional File 2).

Comparative genomics of the chromosome

The chromosomes of the reference strains were aligned and compared using MAUVE, and the overall structure is conserved across all eight chromosomes (Figure S1). In total, 8,348 chromosomal genes were identified in the eight ETEC strains with 3,179 genes considered part of the core genome shared by all eight reference strains. The majority of human commensal *Escherichia coli* (*E. coli*) strains belong to subgroup A [29,30]. However, ETEC strains fall into multiple phylogenetic groups (A, B1, B2, D, E, F and CladeI with the majority found in the phylogenetic groups A and B1 [18]. The phylogenetic group of the eight ETEC reference strains have previously been determined using the triplex-PCR scheme [31]. The ETEC references were re-analysed using ClermonTyping [32] and it was determined that strain E1373 belongs to the

phylogenetic group E while the other reference isolates belong to groups A and B1 (Table 1).

Plasmids

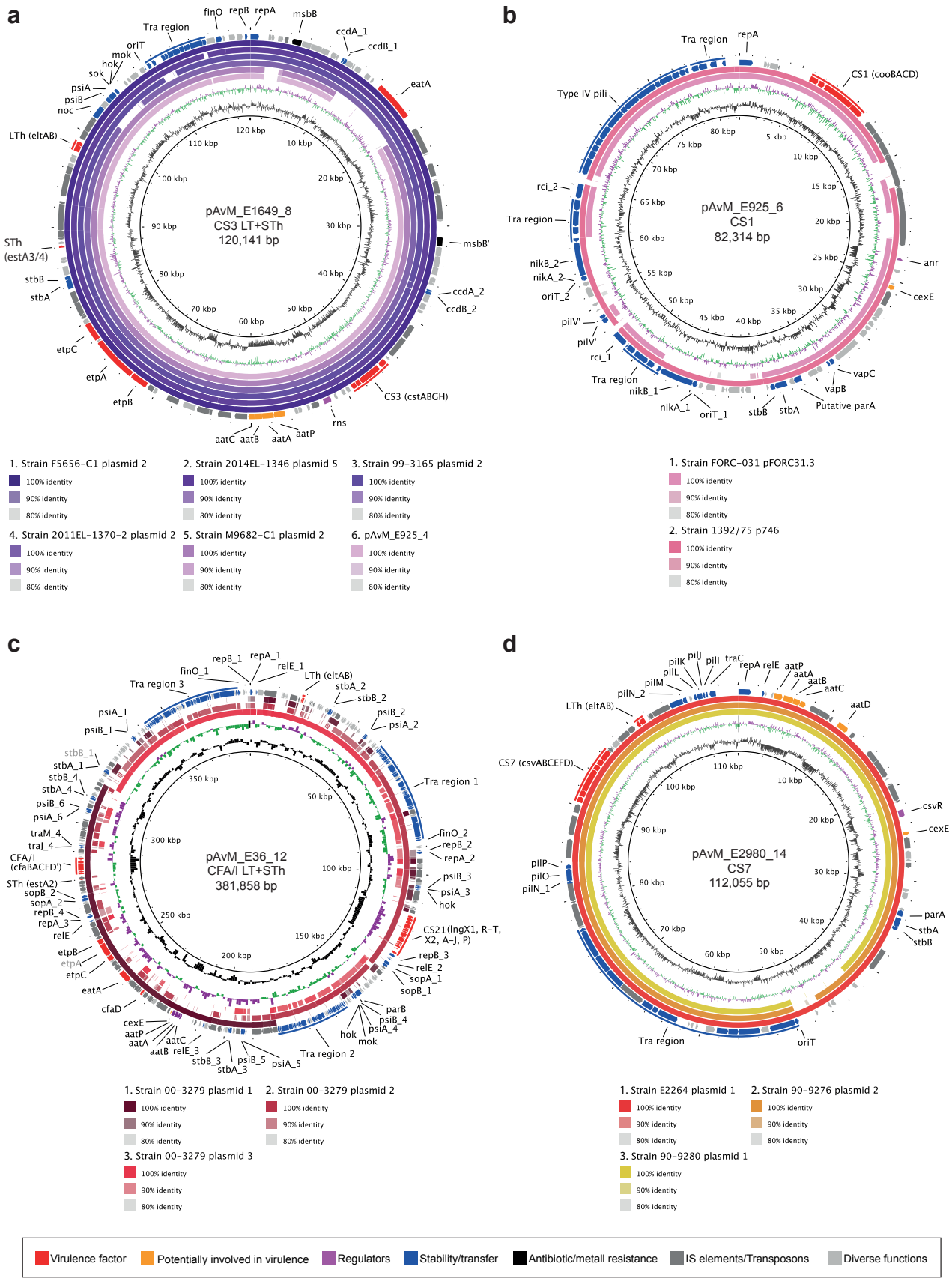
The plasmids of each isolate were annotated using Prokka followed by manual curation of the annotations including genes part of the conjugation machinery and known plasmid stability genes. Virulence factors (including CFs, toxins and putative virulence factors), antibiotic resistance determinants with the Comprehensive Antibiotic Resistance Database (CARD) [33] as well as complete and partial insertion elements and prophages were manually annotated. The plasmids were designated pAvM_strainID_integer, e.g. pAvM_E925_4 (Additional file 2). The first plasmid reported in this study starts at 4 as three previous plasmids E873p1-3 already have been deposited to GenBank related to a different project [8].

Plasmids were typed by analysing the presence and variation of specific replication genes to assign the plasmids to incompatibility (Inc) groups. The Inc groups of the ETEC reference plasmids were first determined using PlasmidFinder and further classified into subtypes using pMLST [34]. The replicons identified are IncFII, IncFIIA, IncFIIS, IncFIB, IncFIC, IncI1 and IncY. Plasmids with replicon IncY, IncFIIY or IncB/O/K/Z mainly harboured plasmid associated genes, such as stability and transfer genes. Importantly, replicons FII, FIB and I1 were found to be strongly associated with virulence genes as all CF loci, toxins and virulence factors *eatA* and *etpBAC* were present on these plasmids. The majority of all ETEC plasmids analysed here (17/29) belong to IncFII, of which six of the IncFII plasmids have an additional IncFIB replicon. In six of the ETEC reference strains two or three IncFII replicons are

present, for example, in strain E925, the plasmids pAvM_E925_4 and 7 both belong to IncFII. However, the plasmids were further subtyped to FII-111 and FII-15, respectively, (Table 2 and Additional file 3), explaining the plasmid compatibility.

Virulence factors

The CFs expressed by the selected reference strains are CFA/I, CS1-CS3, CS5-CS7 and CS21. Three of the strains (E925, E1649 and E1779) express both LT and ST, two strains (E2980 and E1441) express LT and the strains E36 and E562 express STh, while E1373 express STp (Table 1). A plasmid can harbour multiple virulence genes, usually a CF locus and genes encoding one or two toxins. Interestingly, plasmids do not often harbour multiple CF loci, but on individual plasmids (in the ETEC reference strains described here). Exceptions for this is strain E1779 in which CS5 and CS6 loci are located on the same plasmid (pAvM_E1779_19). In both E925 (L1) and E1649 (L2) the genes encoding CS3 (*cstABGH*), ST (*estA*) and LT (*eltAB)* are located on the same plasmid, both with the FII replicon and of roughly the same size (Table 2). Blastn comparison between the plasmids and additional plasmids that harbour the same virulence genes shows that they are highly conserved (Figure 1). The results correspond with the close genetic relationship and common ancestry of lineage 1 (L1) and lineage 2 (L2) [18].


a
pAvM_E1649_8
CS3 LT+STh
120,141 bp
finO
repB
repA
msbB
Tra region
oriT
mok
hok
sok
psiA
psiB
noc
LTh (eltAB)
STh (estA3/4)
stbB
stbA
etpC
etpA
etpB
aatC
aatB
aatA
aatP
rns
CS3 (cstABGH)
ccdA_2
ccdB_2
msbB'
eatA
ccdA_1
ccdB_1
1. Strain F5656-C1 plasmid 2
2. Strain 2014EL-1346 plasmid 5
3. Strain 99-3165 plasmid 2
4. Strain 2011EL-1370-2 plasmid 2
5. Strain M9682-C1 plasmid 2
6. pAvM_E925_4
100% identity
90% identity
80% identity
b
pAvM_E925_6
CS1
82,314 bp
Tra region
repA
CS1 (cooBACD)
Type IV pili
rci_2
nikB_2
nikA_2
oriT_2
pilV'
rci_1
nikB_1
nikA_1
oriT_1
stbB
stbA
Putative parA
vapB
vapC
cexE
anr
1. Strain FORC-031 pFORC31.3
2. Strain 1392/75 p746
c
pAvM_E36_12
CFA/I LT+STh
381,858 bp
repB_1
repA_1
finO_1
relE_1
LTh (eltAB)
Tra region 3
stbA_2
stbB_2
psiA_1
psiB_1
psiB_2
psiA_2
stbB_1
stbA_1
stbB_4
stbA_4
psiB_6
psiA_6
traM_4
traJ_4
CFA/I (cfaBACED')
STh (estA2)
sopB_2
sopA_2
repB_4
repA_3
relE
etpB
etpA
etpC
eatA
cfaD
cexE
aatP
aatA
aatC
aatB
relE_3
stbB_3
stbA_3
psiB_5
psiA_5
Tra region 2
hok
mok
psiA_4
parB
psiB_4
sopB_1
sopA_1
relE_2
repB_3
CS21(lngX1, R-T, X2, A-J, P)
hok
psiA_3
psiB_3
repA_2
repB_2
finO_2
Tra region 1
1. Strain 00-3279 plasmid 1
2. Strain 00-3279 plasmid 2
3. Strain 00-3279 plasmid 3
100% identity
90% identity
80% identity
d
pAvM_E2980_14
CS7
112,055 bp
pilK
pilJ
pilI
traC
repA
relE
aatP
aatA
aatB
aatC
aatD
pilM
pilL
pilN_2
LTh (eltAB)
CS7 (csvABCEFD)
pilP
pilO
pilN_1
Tra region
oriT
parA
stbA
stbB
csvR
cexE
1. Strain E2264 plasmid 1
2. Strain 90-9276 plasmid 2
3. Strain 90-9280 plasmid 1
100% identity
90% identity
80% identity
Virulence factor
Potentially involved in virulence
Regulators
Stability/transfer
Antibiotic/metall resistance
IS elements/Transposons
Diverse functions

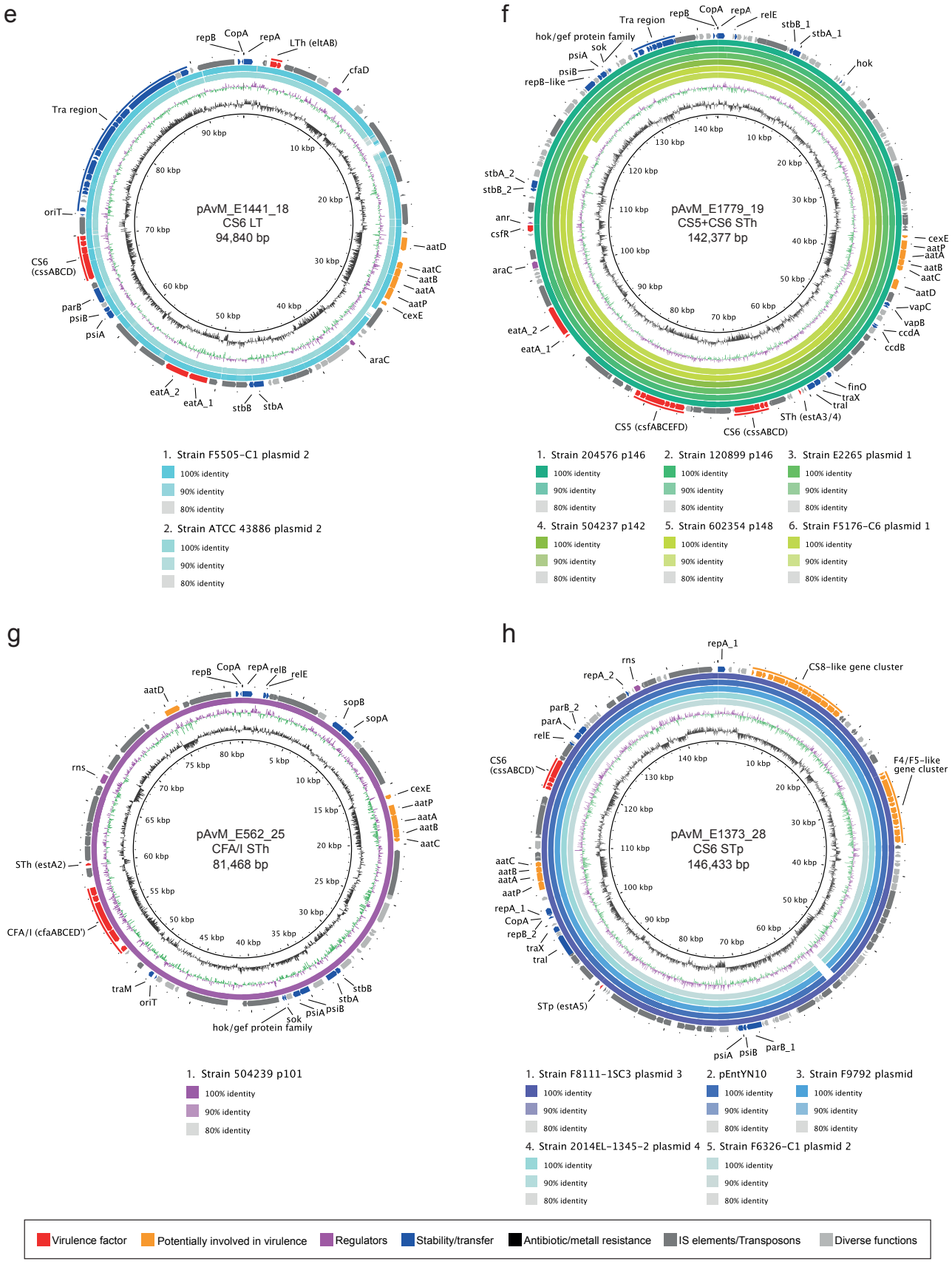

e
repB
CopA
repA
LTh (eltAB)
cfaD
Tra region
oriT
CS6 (cssABCD)
parB
psiB
psiA
eatA_2
eatA_1
stbB
stbA
aatD
aatC
aatB
aatA
aatP
cexE
araC
pAvM_E1441_18
CS6 LT
94,840 bp
1. Strain F5505-C1 plasmid 2
100% identity
90% identity
80% identity
2. Strain ATCC 43886 plasmid 2
100% identity
90% identity
80% identity
f
Tra region
repB
CopA
repA
relE
stbB_1
stbA_1
hok/gef protein family
sok
psiA
psiB
repB-like
hok
stbA_2
stbB_2
anr
csfR
araC
eatA_2
eatA_1
cexE
aatP
aatA
aatB
aatC
aatD
vapC
vapB
ccdA
ccdB
finO
traX
traI
STh (estA3/4)
CS6 (cssABCD)
CS5 (csfABCEFD)
pAvM_E1779_19
CS5+CS6 STh
142,377 bp
1. Strain 204576 p146
2. Strain 120899 p146
3. Strain E2265 plasmid 1
4. Strain 504237 p142
5. Strain 602354 p148
6. Strain F5176-C6 plasmid 1
100% identity
90% identity
80% identity
g
repB
CopA
repA
relB
relE
aatD
sopB
sopA
rns
cexE
aatP
aatA
aatB
aatC
STh (estA2)
CFA/I (cfaABCED')
traM
oriT
sok
hok/gef protein family
psiA
psiB
stbA
stbB
pAvM_E562_25
CFA/I STh
81,468 bp
1. Strain 504239 p101
100% identity
90% identity
80% identity
h
repA_1
rns
repA_2
CS8-like gene cluster
parB_2
parA
relE
CS6 (cssABCD)
F4/F5-like gene cluster
aatC
aatB
aatA
aatP
repA_1
CopA
repB_2
traX
traI
STp (estA5)
psiA
psiB
parB_1
pAvM_E1373_28
CS6 STp
146,433 bp
1. Strain F8111-1SC3 plasmid 3
2. pEntYN10
3. Strain F9792 plasmid
4. Strain 2014EL-1345-2 plasmid 4
5. Strain F6326-C1 plasmid 2
100% identity
90% identity
80% identity
Virulence factor
Potentially involved in virulence
Regulators
Stability/transfer
Antibiotic/metall resistance
IS elements/Transposons
Diverse functions

Figure 1: Comparison between the ETEC reference plasmids harbouring colonisation factors and other PacBio-sequenced ETEC plasmids using blastn.
a) pAvM_E1649_8 (CS3) as reference and pAvM_925_4 (CS3) compared to the following ETEC plasmids: F5656-C1 plasmid 2 (USA, CP024262.1), 2014-EL-1346-6 plasmid 5 (2014, USA; CP024237.1), 99-3165 plasmid 2 (USA; CP029980.1), 2011EL-1370-2 plasmid 2 (2011, USA; CP022914.1) and M9682-C1 plasmid 2 (1975, USA; CP024277.1). **b)** pAvM_E925_6 (CS1) compared to ETEC plasmids pFORC31.3 (2004, Korea; CP013193.1) and 1392/75 p746 (1973; FN822748.1). **c)** pAvM_E36_12 (CFA/I) compared to plasmids 1-3 (p1: CP024294.1; p2 CP024295.1; p3: CP024296.1) from ETEC strain 00-3279 (USA). **d)** pAvM_E2980_14 (CS7) compared to E2264 plasmid 1 (2006, Bangladesh; CP023350.1), 90-9276 plasmid 2 (1988, Bangladesh; CP024298.1) and 90-9280 plasmid 1 (1988, Bangladesh; CP024241.1). **e)** pAvM_E1441_18 (CS6) compared to F5505-C1 plasmid 2 (2013, Sweden; CP023259.1) and ATCC 43886 plasmid 2 (CP024255.1). **f)** pAvM_E1779_19 (CS5+CS6) compared to 204576 p146 (2010, Mali; CP025908.1), 120899 p146 (2012, Gambia; CP025917.1), E2265 plasmid 1 (2006, Bangladesh; CP023347.1), 504237 p142 (2010, India; CP025863.1), 602354 p148 (2009, Bangladesh; CP025848.1) and F5176-C6 plasmid 1 (1997; CP024668.1). **g)** pAvM_E562_25 (CFA/I) compared to p504239_101 (2010, India; CP025860.1). **h)** pAvM_E1373_28 (CS6) compared to F8111-1SC3 plasmid 3 (USA; CP024272.1), pEntYN10 (1991, Japan; AP014654.2), F9792 plasmid (USA; CP023274.1), 2014EL-1345-2 plasmid 4 (2014, USA; CP024227.1) and F6326-C1 plasmid 2 (1998, USA; CP024265.1). The thresholds chosen for the blastn are shown in the key below each plasmid comparison. The colour code for the annotations are listed at the bottom of the figure. The two most inner rings depict GC content in black and GC Skew- in purple and GC Skew+ in green. The figures were generated using BRIG [84] v0.95.

Besides CFs and toxins, additional virulence factors were identified in the majority of the strains (Table 2), with *eatA* and *etpBAC* being the most commonly found.

EatA is an immunogenic mucinase that contributes to virulence by degrading MUC2 which is the major protein component of mucus in the small intestine [35,36]. The *etpABC* genes encode an adhesin located on the tip of the flagella and mediate adherence to host cells [37,38]. Four reference strains (E925, E1649, E36 and E562) harbour both *eatA* and *etpBAC.* In three strains the *eatA* and/or *etpBAC* are located on the same plasmid with an FII or FII+FIB replicon along with additional ETEC virulence genes, except in E562 and E1373, where *eatA* and *etpBA*C are located on an I1+FII

(pAvM_E562_23) and I1 (pAvM_E1373_16) plasmid, respectively, which mainly contains plasmid associated genes including genes encoding the *pil* operon and *tra*-operon (pAvM_E562_23). Furthermore, a less explored putative virulence factor is CexE, which is an extracytoplasmic protein dependent on the expression of the CFA/I regulator *cfaD* [39], and was first identified in H10407 [40]. Corroborating earlier findings, the CFA/I positive E36 (L3) and E562 (L6) isolates harbour *cexE* (pAvM_E36_12 and pAvM_E562_25). In addition, *cexE* is present in pAvM_E925_6, pAvM_E1779_19 and pAVM_E2980_14, pAvM_E1441_18 and pAvM_E1373_28. *CexE* has previously also been identified in several CS5+CS6 positive ETEC and shown to be upregulated in the presence of bile and sodium glycocholate-hydrate [41]. Bile is known to be involved in the regulation of several ETEC CFs [42,43]. The location of *cexE* seems to be conserved across specific strains. In pAvM_E36_12, pAvM_E1441_18, pAvM_E1779_19 and pAvM_E562_25 *cexE* is located upstream of the *aatPABC* locus, whereas in pAvM_E925_6 and pAvM_E2980_14 *cexE* is located downstream of *rob* (an AraC family transcriptional regulator) in the opposite direction. pAvM_E925_4 harbours the *aatPABC* locus; however *cexE* is located on a different plasmid (pAvM_E925_6) in this strain.

Comparison of plasmids with the same virulence profile

ETEC isolates within a lineage share the same virulence profile, specifically the same CF profile (Figure S10). We verified that our selected isolates grouped within previously described lineages with confirmed virulence profiles by phylogenetic analyses (Figure S2). Blastn of each of the CF positive plasmids from each reference genome were performed, and the best hit(s) were used for subsequent analysis (Figure 1). Most of the plasmids identified as related to the ETEC reference plasmids were not

annotated, hence, when needed these were annotated using the corresponding ETEC reference plasmids annotation as a high priority when running Prokka. We show that plasmids with the same CF and toxin profile from the same lineage are often conserved (Figure 1). For example, the two plasmids encoding CS3 (pAvM_E925_4 and pAvM_E1649_8) are highly similar to several CS3 harbouring plasmids from O6:H16 strains collected from various geographical locations between 1975 and 2014, including *E. coli* O6:H16 strain M9682-C1 plasmid unnamed2 (CP024277.1) and *E. coli* strain O6:H16 F5656C1 plasmid unnamed2 (CP024262.1) PacBio sequenced by Smith et al. [20] (Figure 1a). Furthermore, high coverage and similarity were found between the plasmids of isolates E1441 (L4), and PacBio sequenced plasmids of ETEC isolates ATCC 43886/E2539C1 and 2014EL-1346-6 [20]. These isolates were collected in the seventies [44] and 2014 (from a CDC collection), respectively, and assigned as O25: H16 which is the O group determined for E1441 in silico (Figure 1e). Plasmids of E2980 (LT+ CS7, L3) were validated by the PacBio sequenced plasmids of ETEC isolate E2264 (Figure 1d). Similarly, two plasmids of E1779 (LT, STh+CS5+CS6, L5) was identified in E2265 (LT, STh+CS5+CS6 [28,41], although E1779 harboured two additional plasmids. Several additional L5 ETEC genomes have been sequenced within the GEMS study [45], and high plasmid similarity and conservation in CS5+CS6 positive L5 isolates was evident (Figure 1f).

Overall the results show that ETEC plasmids are specific to lineages circulating worldwide and conserved over time (Figure 1, Figure S2-S9 for more extensive plasmid annotation and Figure S10). Thus, the plasmids of major ETEC lineages must confer evolutionary advantages to their host genomes since they are seldom lost.

Antibiotic resistance

E. coli can become resistant to antibiotics, both via the presence of antibiotic resistance genes and the acquisition of adaptive and mutational changes in genes encoding efflux pumps and porins which allows the bacterium to pump out the antibiotic molecules effectively.

Antibiotic resistance genomic marker(s), both chromosomally located and on plasmids, were identified using the CARD database [33] (Table 2, Figure S12 and Figure S13 and Additional file 2). Similar to other studies, IncFII and B/O/K/Z plasmids were found to harbour genes conferring antibiotic resistance [46]. Furthermore, the phenotypic antibiotic resistance profile was determined with clinical MIC breakpoints based on EUCAST (The European Committee on Antimicrobial Susceptibility Testing) [47] (Table S2). Phenotypic antibiotic resistance profiles (Table S2) were supported mainly by the findings of antibiotic resistance genes, efflux pumps and porins (Figures S4 and S5 and Table S3), although some differences were found. All ETEC reference strains are phenotypically resistant to at least two antibiotics of the 14 tested (Table S2). Resistance against penicillin's, norfloxacin (Nor) and chloramphenicol (Cm) is most common among these strains. Two of the strains, E1441 and E2980, harbour more than four antibiotic resistance genes as well as multiple efflux systems and porins (Figure S12, Figure S13 and Table S3). The plasmid pAvM_E1441_17 carries *aadA1-like*, *dfrA15*, *sul1* and *tetA*(A) resistance genes (Table 2), where the first three genes are in a Class 1 integron which confers resistance to streptomycin, trimethoprim, and sulphonamide (sulphamethoxazole). The gene *tetA*(A) is part of a truncated Tn*1721* transposon [48]. The E1441 strain was verified as resistant to tetracycline (Tet) and sulphamethoxazole-trimethoprim (Sxt) while streptomycin was not tested. A *mer*

operon derived from Tn*21* is also present the resistance region of pAvM_E1441_17 (Table 2), indicating that the plasmid would also likely confer tolerance to mercury, although this was not confirmed. Interestingly, this multi-replicon (FII and FIB) plasmid also harbours the *lng* locus encoding CS21, one of the most prevalent ETEC CFs. In isolate E2980 virulence plasmid pAvM_E2980_15 harboured multiple resistance genes in the same region (*bla*$_{\text{TEM-1b}}$, *strA*, *strB* and *sul2)* conferring resistance to ampicillin, streptomycin and sulphonamides. E2980 was found to be resistant to ampicillin (Amp) and oxacillin (Oxa), which can be broken down by the beta-lactamase Bla$_{\text{TEM-1b}}$, (Table 2, Table S2 and Table S3). E562 harbours three antibiotic resistance genes, *ampC* located in the chromosome and the *tet*(A) and *bla*$_{\text{TEM-1b}}$ genes on an FII plasmid (pAvM_E562_27). The *mer* operon derived from Tn*21* is also present in the region (Table 2 and Table S3). The phenotypic resistance profile of E562 matches the genomic profile with resistance to tetracycline (Tet), ampicillin (Amp), amoxicillin-clavulanic acid (Amc) and oxacillin (Oxa) (Table S2). The plasmid pAvM_E36_13 contains a complete copy of Tn*10*, which encodes the *tet*(B), tetracycline resistance module. Although the AvM_E1373_29 phage-like plasmid is cryptic, related plasmids such as the pHMC2-family of phage-like plasmids [49] (described below), can harbour resistance genes such as *bla*$_{\text{CTX-M-1}}$4 [50] and *bla*$_{\text{CTX-M-15}}$ [51,52].

Phenotypic intermediate resistance to ampicillin was found in E36 and E1779 encoded by chromosomal *ampC*. Higher MIC values against ampicillin are found in E2980 and E562 strains carrying *bla*$_{\text{TEM}}$ genes. Phenotypic resistance to ceftazidime (Caz) and ceftriaxone (Cro) was not found in the isolates, which were consistent with the absence of extended-spectrum beta-lactamase (ESBL) resistance genes in the sequence data.

312 Resistance to chloramphenicol (Cm) was found in five isolates, but none of the resistant
313 isolates contained known resistance genes suggesting that chromosomal mutations or
314 presence of efflux pumps may account for this reduced susceptibility.
315 The ETEC reference strains contain several efflux systems which could explain why
316 the genotypic and phenotypic antibiotic resistance profile did not match for all
317 antibiotics. All of the isolates harbour multiple efflux pumps located on the
318 chromosome and plasmids (Table S3 and Figure S12). In E925, a non-synonymous
319 mutation in *acrF* was identified (G1979A) resulting in a substitution from arginine to
320 glutamine (A360Q). The effect on the expression and/or function of the AcrEF efflux
321 pump was not verified.
322
323 Phenotypic resistance to norfloxacin (Nor) was found in 6 of the isolates. The isolates
324 were analysed for chromosomal mutations likely to confer quinolone resistance, using
325 ResFinder but mutations in *gyrA* were only found in one strain, E2980, at position S83A
326 which may confer resistance to nalidixic acid, norfloxacin and ciprofloxacin. However,
327 E2980 was sensitive to Nalidixic acid. Both mutation(s) that alter the target (*gyrA* and
328 *parC*), as well as the presence of efflux pumps, can confer resistance to
329 fluoroquinolones. The majority of the isolates are moderately resistant to Norfloxacin
330 (and Nalidixic acid), both quinolones, which is most likely due to the presence of two
331 efflux pumps, AcrAB-R and AcrEF-R, as only one mutation was identified in *gyrA* of
332 isolate E2980 where usually at least two or more mutations are needed to confer
333 augmented resistance [53].
334
335 ***Identification of phage-like plasmids in ETEC***

Two of the ETEC reference strains (E1649 and E1373) harboured phage-like plasmids (pAvM_E1649_9 and pAvM_E1373_29) which encode for DNA metabolism, DNA biosynthesis as well as structural bacteriophage genes (capsid, tail etc.). Both pAvM_E1649_9 and pAvM_E1373_29 contain genes associated with plasmid replication, division and maintenance (i.e. *repA* and *parAB*). Phage-like plasmids are found in various bacterial species, such as *E. coli*, *Klebsiella pneumoniae*, *Yersinia pestis*, *Salmonella enterica* serovar Typhi, *Salmonella enterica* serovar Typhimurium, *Salmonella enterica* serovar Derby and *Acinetobacter baumanii* [54]. The plasmid pAvM_E1649_9 belong to the P1 phage-like plasmid family (Figure 2a and Figure S14a) while pAvM_E1373_29 belongs to the pHCM2-family (Figure 2b and Figure S14b) that can be traced back to a likely phage origin similar to the *Salmonella* phage, SSU5 [49]. Both phage-plasmids thus contain replication and/or partition genes of plasmid origin and a complete set of genes that are phage related in function and properties (Figure 2 and Figure S14). Significantly, phage-like plasmid pAvM_E1373_29 falls more within the *E. coli* lineage of pHCM2 phage-like plasmid rather than those found in *Salmonella* species. This indicates that phage-like plasmids have diversified within the bacterial species they were isolated.

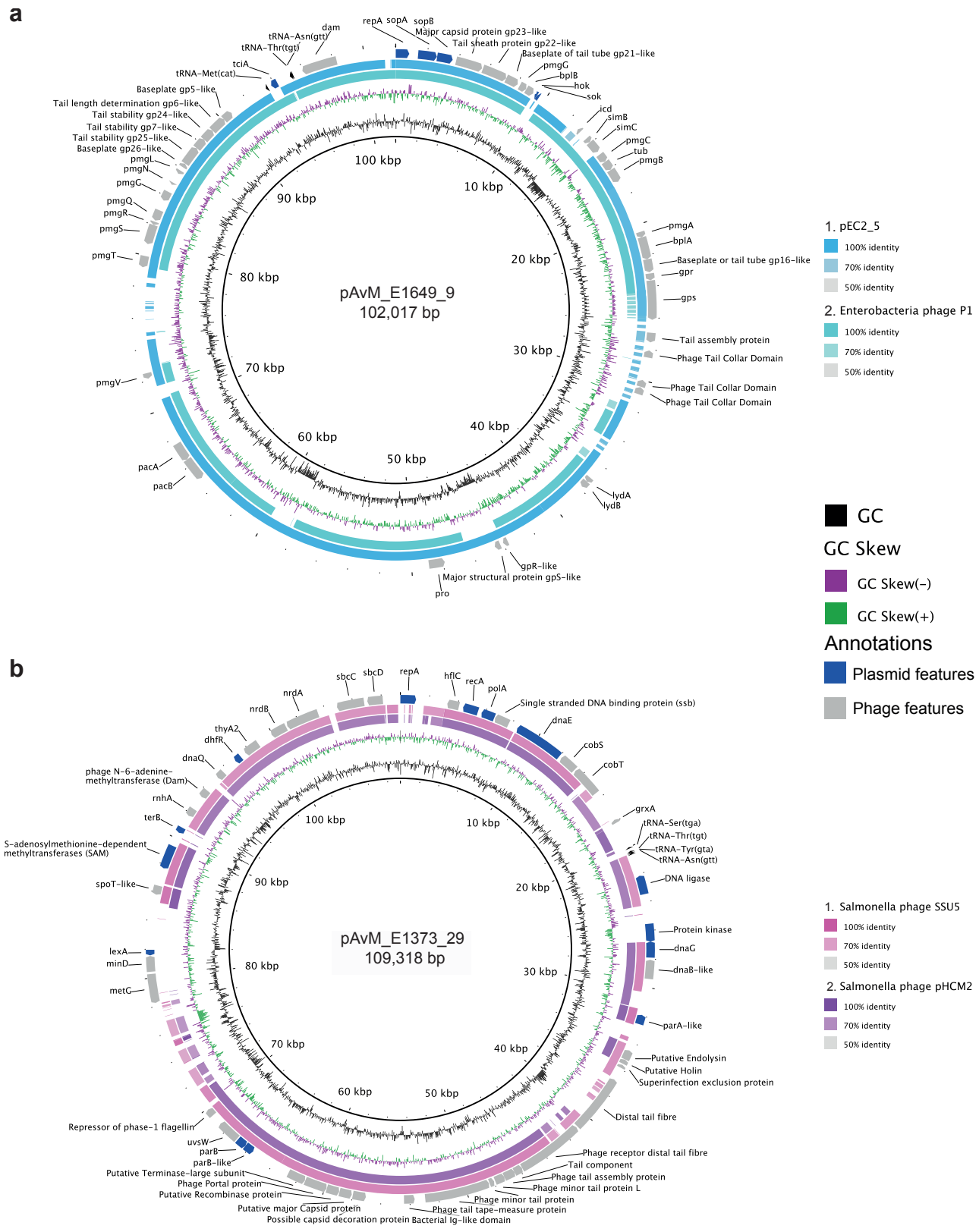

a
pAvM_E1649_9
102,017 bp
100 kbp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
repA
sopA
sopB
Major capsid protein gp23-like
Tail sheath protein gp22-like
Baseplate of tail tube gp21-like
pmgG
bplB
hok
sok
icd
simB
simC
pmgC
tub
pmgB
pmgA
bplA
Baseplate or tail tube gp16-like
gpr
gps
Tail assembly protein
Phage Tail Collar Domain
Phage Tail Collar Domain
Phage Tail Collar Domain
lydA
lydB
gpR-like
Major structural protein gpS-like
pro
pacB
pacA
pmgV
pmgT
pmgS
pmgR
pmgQ
pmgG
pmgN
pmgL
Baseplate gp26-like
Tail stability gp25-like
Tail stability gp7-like
Tail stability gp24-like
Tail length determination gp6-like
Baseplate gp5-like
tRNA-Met(cat)
tciA
tRNA-Thr(tgt)
tRNA-Asn(gtt)
dam
1. pEC2_5
100% identity
70% identity
50% identity
2. Enterobacteria phage P1
100% identity
70% identity
50% identity
GC
GC Skew
GC Skew(-)
GC Skew(+)
Annotations
Plasmid features
Phage features
b
pAvM_E1373_29
109,318 bp
100 kbp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
sbcC
sbcD
repA
hflC
recA
polA
Single stranded DNA binding protein (ssb)
dnaE
cobS
cobT
grxA
tRNA-Ser(tga)
tRNA-Thr(tgt)
tRNA-Tyr(gta)
tRNA-Asn(gtt)
DNA ligase
Protein kinase
dnaG
dnaB-like
parA-like
Putative Endolysin
Putative Holin
Superinfection exclusion protein
Distal tail fibre
Phage receptor distal tail fibre
Tail component
Phage tail assembly protein
Phage minor tail protein L
Phage minor tail protein
Phage tail tape-measure protein
Bacterial Ig-like domain
Possible capsid decoration protein
Putative major Capsid protein
Putative Recombinase protein
Phage Portal protein
Putative Terminase-large subunit
parB-like
parB
uvsW
Repressor of phase-1 flagellin
metG
minD
lexA
spoT-like
S-adenosylmethionine-dependent methyltransferases (SAM)
phage N-6-adenine-methyltransferase (Dam)
rnhA
terB
dnaQ
dhfR
thyA2
nrdB
nrdA
1. Salmonella phage SSU5
100% identity
70% identity
50% identity
2. Salmonella phage pHCM2
100% identity
70% identity
50% identity

Figure 2: Comparisons between the identified ETEC phage-plasmids and other similar phage-plasmids using blastn. **a)** pAvM_E1649_9 is a P1-like phage-plasmid here compared to Enterobacteria phage P1 (Escherichia virus P1; NC_005856.1) and pEC2_5 (*E. coli* strain EC2_5; CP041960.1). **b)** pAvM_E2980_29, a phage-plasmid similar to the pHCM2 (Salmonella typhi strain CT18; AL513384.1) and SSU5 (Salmonella phage; JQ965645.1) in Salmonella typhi. Blastn comparisons were made using BRIG [84] (v0.95) with the thresholds indicated to the right of each plasmid comparison. Selected phage and plasmid annotations are shown in the outer ring.

Blastn searches confirmed high similarity (at least 80% at the DNA across much of the sequence) of pAvM_E1373_29 to several phage-like plasmids found in *E. coli* including ETEC O169:H41 isolate F8111-1SC3 [20,55], several $bla_{\text{CTX-M-15}}$ positive phage-like plasmids (pANCO1, pANCO2 [52] and PV234a), as well as a plasmid found in *E. coli* ST648 from wastewater and ST131 isolate SC367ECC [56]. The P1 phage-like plasmid pAvM_1649_9 is most similar to p1107-99K, pEC2_5 isolated from human urine and p2448-3 from a UPEC ST131 isolate isolated from blood. The similarity is most pronounced at the amino acid level. Conservation and synteny are evident when pAvM_1649_9 is compared to P1 phage.

Prophages present and their cargo genes

Prophages may insert into chromosomes and bring along genes required for lysogeny and lytic cycles and cargo genes that are often picked up when DNA is compacted into the capsid. Cargo genes can significantly benefit the host bacterium by providing additional elements to defence against phage or immune evasion and finally, environmental survival. PHASTER analyses identified prophages in the chromosomes of all ETEC reference isolates and some of the plasmids (Table S4). Putative tellurite resistance operons in isolates E925, E36, E2980 and E1373 were all located in

389 prophages. In addition, *eatA* (in E925, and E1649) and *estA* (STh) genes (E36) were
390 prophage cargo genes.
391
392 Many prophage cargo genes identified in this study have properties related to inhibition
393 of cell division. Among these are a variety of *kil* genes which can enhance host bacterial
394 survival in the presence of some antibiotics [57]. Some genes that are core entities
395 within many prophages, such as *zapA* (from E1779_Pph_6), *dicB* and *dicC* (found in
396 phage E1779_Pph_7), also have similar effects as they can inhibit cell division in the
397 presence of antibiotics which raise the broader question in terms of how they are
398 beneficial to the host bacterium.
399
400 A different gene of interest is the *yfdR* gene identified in E1779_Pph_7 (gene
401 E1779_04412). YfdR curtails cellular division by inhibiting DNA replication under
402 stress conditions encountered by the bacterial cell. Similarly, the *iraM* gene located in
403 phage E1441_Pph_2 plays a role in RpoS stability.
404
405 OmpX homologs were found in numerous phages in this study. They are trans-
406 membrane located and play a role in virulence as well as antibiotic resistance [58]. PerC
407 is often associated with EPEC plasmids, where it seems to have a regulatory role for
408 the attaching and effacing gene, *eaeA* [59]. Its presence as cargo within an ETEC strain
409 phage, E1779_Pph_7 located on the chromosome, is intriguing. Its ability to regulate
410 other virulence genes is yet to be determined. Within the same phage, a *gntR*-like
411 regulatory gene was identified. This gene plays a role in gluconate utilisation and
412 induction of the Entner-Doudoroff pathway [60].
413

Discussion

ETEC strains have previously been shown to fall into globally spread genetically conserved lineages which encompass strains with specific virulence factor profiles [18]. The currently widely used ETEC reference strains H10407 (CFA/I) and E24377A (CS1+CS3) are highly divergent from other strains with the same virulence profile sequenced more recently [18] and highlights the need for relevant and representative ETEC reference strains and genomes. The long-read sequenced isolates presented here comprise complete reference genomes with separate chromosomal and plasmid sequences that allow more detailed studies of ETEC and *E. coli* phylogeny. The reference strains are representative isolates of their respective lineage and cluster phylogenetically together with different ETEC isolates sequenced by several other groups (Figure S10).

Previous studies confirmed that ETEC belongs to lineages that have spread globally. These analyses were mainly dependent on the shared core genome of chromosomal genes while conservation of plasmids was indicated by the association between the plasmid-borne toxin and CFs and lineage [18]. Analysis of the plasmids sequenced in the present study showed that the conservation within ETEC lineages also include plasmids.

Blast analyses confirm that the plasmids identified in this study are often highly homologous to other plasmids identified by either long-read or Illumina sequencing present in GenBank. For instance, the 94.5 kb plasmid pAvM_1441_18 was 98% identical to two 96 kb and 82 kb plasmids belonging to ETEC O25: H16 isolates ATCC 43886/E2539C1 and 2014EL-1346-6 sequenced by PacBio by Smith *et al.,* [20] (Figure 1e and Figure S6). Plasmid pAvM_E1441_18 is the major virulence plasmid of this lineage carrying genes encoding LT and CS6.

The larger plasmid in E1441 (pAvM_E1441_17) carries both the genes for ETEC CF CS21 and antibiotic resistance determinants. Furthermore, complete conjugation machinery was present suggesting that this is most likely a self-transmissible plasmid, though this was not confirmed. Movement of such a plasmid would result in the spread of ETEC virulence genes and AMR determinants.

Interestingly, Wachsmuth et al., [44] analysed transfer frequencies in ETEC O25:H16 isolates (the same serogroup was identified in E1441) and found evidence that resistance to tetracycline and sulfathiazole was transferred but not the genes encoding LT [44]. The same study found evidence of two large plasmids of similar size [44] corroborating our findings of two plasmids of similar size in E1441, one with *eltAB* and *cssABCD* without the *tra*-operon (pAvM_E1441_18) and the other putatively mobile plasmid (pAvM_E1441_17) carrying the *sul1* and *tet*(A) genes as well as the *lng* operon encoding CF CS21. Since ATCC 43886/E2539C1, E1441 and 2014EL-1346-6, have been isolated in the 1970s, 1997, and 2014, respectively, our findings indicate that E1441 represent an ETEC lineage with stable plasmid content and putative ability to transfer antibiotic resistance and the CS21 operon by transfer of one of the plasmids. Furthermore, pAvM_E1441_17 is a multi-replicon plasmid. Multi-replicon plasmids have been described as a way to broaden their host range, i.e. possibility to be transferred between bacteria of different phylogenetic groups [61,62]. Whether this plasmid type is found in other *E. coli* remains to be investigated but the finding that the L4 lineage retains both plasmids in isolates collected over time and worldwide indicate a strong selective force to keep the extra-chromosomal contents of both plasmids.

The ETEC O169:H41 isolate F8111-1SC3 plasmid unnamed 2 [20,55] is highly similar to pAvM_E1373_28 (Figure 1h and Figure S9). The F8111-ISC3 isolate is part of a CDC collection of ETEC isolates from cruise ship outbreaks and diarrheal cases in US

1996-2003. The antibiotic resistance profiles of these isolates were determined [55] and most isolates of O group 169 were tetracycline resistant consistent with the findings of the *tet* gene in E1373 isolated in Indonesia in 1996. ETEC diarrhoea caused by O169:H41 and STp CS6 isolates is repeatedly reported to cause diarrhoea, particularly in Latin America [45,63–65]. Among the cruise ship isolates is the sequenced and characterised virulence plasmid pEntYN10 encoding STp and CS6, described as unstable and easily lost in vitro [63,66]. The E1373 plasmid; AvM_E1373_28 is highly homologous to pEntYN10 (Figure 1h and Figure S9) and the virulence profile of ETEC O169: H41 is conserved in isolates collected globally. Hence, the instability of the plasmid is incongruent with current data indicating that plasmids are stable within this lineage and serotype.

Interestingly, two distinctive extra-chromosomal elements which are highly similar to P1 and SSU5 phage were identified among the 8 ETEC reference strains sequenced (Figure 2, Figure S14 and Table S4). The SSU5-like element carries several genes that allow it to be functional as a plasmid and belongs to the pHCM2-like family of Phage-Plasmids (Figure 2b) [49]. These plasmids are devoid of virulence factors, transposons and antibiotic markers but, they contain a significant number of DNA metabolism and biosynthesis genes and they may contain bacteriophage inhibitory genes that have not yet been identified. Interestingly, several SSU5 phage-like plasmids have been shown to carry the ESBL gene $bla_{\text{CTX-M15}}$ in extra-intestinal pathogenic *E. coli* isolates [51]. ESBL resistance seems to be absent or low in ETEC and the SSU5 phage-like plasmid pAvM_E1373_29 does not contain antibiotic resistance genes. A recent study investigating the distribution of phage-plasmids show that the phage homologs tend to be more conserved and the plasmid homologs more variable [67]. This is also seen in

the phage-plasmids identified here, e.g., genes that could be advantageous to the host cell linked to metabolism and biosynthesis.

To summarise, we provide fully assembled chromosomes and plasmids with manually curated annotations that will serve as new ETEC reference genomes. The in-depth analysis of gene content, synteny and correct annotations of plasmids will also help to elucidate other plasmids with and without virulence factors in related bacterial species. The ETEC reference genomes compared to other long-read sequenced ETEC genomes confirm that the major ETEC lineages harbour conserved plasmids that have been associated with their respective background genomes for decades. This confirms that the plasmids and chromosomes of ETEC are both crucial for ETEC virulence and success as pathogens.

Methods

Selection of strains

Initially one to two ETEC strains within each of the lineage (L1-L7)-specific CF profile were chosen from the University of Gothenburg large collection of ETEC strains [18] for PacBio sequencing. The strains were selected based on the location and year of isolation to represent strains isolated from patients with diarrhoea from diverse geographical locations and at different time-points. After the genomes had been sequenced, assembled, circularised and annotated a second selection was made for manual curation of the genomes. This selection was made based on the quality of the genome assembly and the circularisation. The whole genomes of the ETEC reference strains were compared with one or two other long-read sequenced ETEC strains belonging to the same lineage by progressiveMAUVE and showed that the strains are colinear (Figure S15). One representative ETEC genome from each lineage was

513 annotated, with emphasis on the plasmids. The physical ETEC reference strains are
514 available upon request.
515
516 ***Phenotypic toxin and CF analyses***
517 ETEC isolates were identified by culture on MacConkey agar followed by an analysis
518 of LT and ST toxin expression using GM1 ELISAs [43]. The expression of the different
519 CFs was confirmed by dot-blot analysis [43]. Isolates had been kept in glycerol stocks
520 at −70 °C, and each strain has been passaged as few times as possible.
521
522 ***Antibiotic susceptibility testing***
523 All ETEC isolates were tested against 14 antimicrobial agents and their minimum
524 inhibitory concentration was determined by broth microdilution using EUCAST
525 methodology [47]. The antimicrobial agents were: ampicillin, amoxicillin-clavulanic,
526 oxacillin, ceftazidime, ceftriaxone, doxycycline, tetracycline, nalidixic acid,
527 norfloxacin, azithromycin, erythromycin, chloramphenicol, nitrofurantoin and
528 sulfamethoxazole-trimethoprim. All antibiotics were purchased from Sigma-Aldrich.
529 The *E. coli* ATCC 25922 was used as quality control. The MIC was recorded visually
530 as the lowest concentration of antibiotic that completely inhibits growth.
531
532 ***DNA extraction and sequencing***
533 Strains from each lineage (L1-L7) were SMRT-sequenced on the PacBio RSII. A
534 hybrid *de novo* assembly was performed combining the reads from both the SMRT-
535 sequenced and Illumina sequenced strains.

536 For Single-Molecule Real-Time (SMRT) sequencing (Pacific Bioscience) long intact
537 strands of DNA are required. The genomic DNA extraction was performed as follows.

Isolates were cultured in CFA broth overnight at 37°C followed by cell suspension in TE buffer (10 mM Tris and 1 mM EDTA pH 8.0) with 25% sucrose (Sigma) followed by lysis using 10 mg/ml lysozyme (in 0.25 Tris pH 8.0) (Roche). Cell membranes were digested with Proteinase K (Roche) and Sarkosyl NL-30 (Sigma) in the presence of EDTA. RNase A (Roche) was added to remove RNA molecules. A phenol-chloroform extraction was performed using a mixture of Phenol:Chloroform:Isoamyl Alcohol (25:24:1) (Sigma) in phase lock tubes (5prime). To precipitate the DNA 2.5 volumes 99% ethanol and 0.1 volume 3 M NaAc pH 5.2 was used followed by re-hydration in 10 mM Tris pH 8.0. DNA concentration was measured using NanoDrop spectrophotometer (NanoDrop). On average 10 μg for PacBio sequencing. Library preparation for SMRT sequencing was prepared according to the manufacturers' (Pacific Biosciences) protocol. The DNA was stored in E buffer and sequenced at the Wellcome Sanger Institute. Isolates were sequenced with a single SMRTcell using the P6-C4 chemistry, to a target coverage of 40–60X using the PacBio RSII sequencer.

Assembly

The resulting raw sequencing data from SMRT sequencing were *de novo* assembled using the PacBio SMRT analysis pipeline (https://github.com/PacificBiosciences/SMRT-Analysis) (v2.3.0) utilising the Hierarchical Genome Assembly Process (HGAP) [68]. For all samples, the unfinished assembly produced a single, non-circular, chromosome plus some small contigs, some of which were plasmids or unresolved assembly variants. Using Circlator [69] (v1.1.0), small self-contained contigs in the unfinished assembly were identified and removed, with the remaining contigs circularised. Quiver [68] was then used to correct errors in the circularised region by mapping corrected reads back to the circularised assembly. As the strains had also been short read sequenced, and this data is of higher base quality,

563 the short reads from the Illumina sequencing were used in combination with the long
564 reads using Unicycler [70] to generate high-quality assemblies.
565
566 Fully circularised chromosomes and plasmids were achieved for the majority of the
567 strains. Cross-validation of the assemblies was performed where two or three strains of
568 a lineage were sequenced (Figure S15). A single assembly from each lineage was
569 chosen to act as the representative reference genome, with priority given to assemblies
570 with the most complete and circularised chromosome and plasmids. In total, one
571 chromosome and 5 out of the 29 plasmids could not be circularised (independent on the
572 two strains that were sequenced initially) out of the 8 selected representative strains.
573 These are indicated in Table 2 and Table S1. Between two and five plasmids were
574 identified in the eight strains. Shorter contigs that could not be assembled properly
575 contained phage genes and are included in the genomes and annotated as prophages
576 Table S4). Socru was used to validate the assembly of the chromosome, they all have
577 biologically valid orientation and order of rRNA operons with a type GS1.0, which is
578 seen in most *E. coli* in the public domain [71].
579
580 ***Phylogenetic tree***
581 The phylogenetic relationship between the ETEC reference genomes to other ETEC
582 and *E. coli* commensals and pathotypes was investigated. The following collections
583 were included: ETEC-362 [18], ECOR [72] and the Horesh collection [73] along with
584 additional ETEC genomes from several studies [20,24,26,27,45,74,75]. The reads of
585 identified ETEC genomes from other studies were downloaded from GenBank and
586 assembled using Velvet. Long-read sequenced ETEC genomes were included in the
587 tree and were not re-assembled. The phylogroup of the ETEC strains was determined

using ClermonTyping [32] (v20.03). The virulence profile of the ETEC strains was determined using ARIBA [76] (v2.14.16) with default settings using the custom ETEC virulence database (https://github.com/avonm/ETEC_vir_db). A total of 1,066 genomes was included in the phylogenetic tree. The alignment of core genes (n = 2,895) identified by Roary [77] (v3.12.0) was converted to a SNP-only alignment using snp-sites [78]. A phylogenetic tree was produced with IQ-TREE [79] (v1.6.10) using a GTR gamma model (GTR+F+I) optimised using the built-in model test and visualised using the R package ggtree [80].

Gene prediction, annotation and comparative analysis

The final assembly was annotated using Prokka [81] (v1.14.6). The annotations of all plasmids generated by Prokka were manually checked using the genome viewer Artemis [82] and Geneious 2018.11.1.5 (http://www.geneious.com) together with blastp. Annotations of known ETEC virulence genes (colonisation factors, toxins, *eatA* and *etpBAC*) were added after blast+ [83] analysis using the reference genes available in the ETEC virulence database (https://github.com/avonm/ETEC_vir_db) and their annotations updated accordingly. The LT and ST alleles were determined according to Joffre et al., (https://github.com/avonm/ETEC_toxin_variants_db) [15,17]. Where required, PFAM domains were searched using jackhammer to back up any identified protein using blastp (https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer). Blastn and tblastx were used for plasmid comparison, using both NCBI website or within BLAST Ring Image Generator (BRIG) [84] (v0.95).

Incompatibility groups

612 Due to the discrepancy in databases two approaches was used to determine the Inc
613 groups of the 25 plasmids. PlasmidFinder was used with a threshold for minimum %
614 identity at 95% and minimum coverage of 60%. The plasmids were further
615 characterised by pMLST [34], except for IncY which are a group of prophages that
616 replicate in a similar manner as autonomous plasmids (Additional File 3). IncB/O/K/Z
617 plasmids were further typed by blastn comparison to the reference B/O (M93062), K
618 (M93063) and Z (M93064) replicons.
619
620 ***oriT prediction***
621 The location of the *oriT* in the plasmids, if present, was predicted using oriTFinder [85]
622 with Blast E-value cut-off set to 0.01.
623
624 ***Genomic antibiotic resistance profiling***
625 The identification of antibiotic resistance genes, located on both the chromosome and
626 plasmid(s) as well as the presence of efflux pumps and porins known to confer
627 resistance to antibiotics. The results were obtained by running ARIBA [76] using the
628 CARD database [86] with the default settings (minimum 90% sequence identity and no
629 length cut-off). ARIBA combines a mapping/alignment and targeted local assembly
630 approach to identify AMR genes and variants efficiently and accurately from paired
631 sequencing reads. The heatmaps were visualised using Phandango [87]. The presence
632 of chromosomal mutations in *gyrA* and *parC* was determined with ResFinder (v3.2)
633 from the Center of Genomic Epidemiology [88].
634
635 ***Virulence gene prediction***

The ETEC assemblies from the ETEC-NCBI collection (Additional file 4) were screened using abricate [89] with default settings against the ETEC virulence database (https://github.com/avonm/ETEC_vir_db) for virulence gene (including *eatA* and *etpBAC*) predication. A subset of the isolates in the ETEC-NCBI dataset have previously been analysed for the presence of EatA where a sample with negative PCR but positive western blots were included as positive [74]. Here, only isolates harbouring the *eatA* and *etpBAC* genes are considered positive.

Prophage prediction

The complete FASTA sequence of each ETEC reference genome was searched for phage genes and prophages using PHASTER (phaster.ca) [90]. The identified intact prophages are listed in Table S4. All prophage contained cargo genes but only recognisable genes are stated, not any hypothetical. Additional questionable and not intact prophages were identified but have not been included here. The prophages have been given a specific identifier name and are also annotated as a mobile_element in the submitted chromosome and or plasmid(s) of each strain.

Insertion sequences

Insertion sequences in the plasmids as well as surrounding the CS2 loci located on the chromosome of E1649 were annotated using both Galileo AMR software [91] and the ISFinder database [92]. Complete and partial IS elements were annotated (>95% identity with hits in ISFinder) along with the present genes encoding transposases. Three new insertion sequences were detected in this analysis and were submitted to ISFinder as TnEc2, TnEc3 and TnEc4. Transposons and other mobile elements

(integrons and group II introns) were also identified using Galileo AMR and blastn against public databases.

References

1. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990-2016. The Lancet Infectious diseases. 2018;18:1229–40.

2. Baron S, Evans DJ, Evans DG. Escherichia Coli in Diarrheal Disease. 1996;

3. Svennerholm A, Lundgren A. Recent progress toward an enterotoxigenic Escherichia coli vaccine. Expert Rev Vaccines. 2014;11:495–507.

4. Svennerholm A-M, Lundgren A. Recent progress toward an enterotoxigenic Escherichia coli vaccine. Expert Review of Vaccines. 2012;11:495–507.

5. Lundgren A, Bourgeois L, Carlin N, Clements J, Gustafsson B, Hartford M, et al. Safety and immunogenicity of an improved oral inactivated multivalent enterotoxigenic Escherichia coli (ETEC) vaccine administered alone and together with dmLT adjuvant in a double-blind, randomized, placebo-controlled Phase I study. Vaccine. 2014;32:7077–84.

6. Harro C, Bourgeois AL, Sack D, Walker R, DeNearing B, Brubaker J, et al. Live attenuated enterotoxigenic Escherichia coli (ETEC) vaccine with dmLT adjuvant protects human volunteers against virulent experimental ETEC challenge. Vaccine. 2019;37:1978–86.

7. Qadri F, Svennerholm A-M, Faruque AS, Sack RB. Enterotoxigenic Escherichia coli in developing countries: epidemiology, microbiology, clinical features, treatment, and prevention. Clinical microbiology reviews. 2005;18:465–83.

8. von Mentzer A, Tobias J, Wiklund G, Nordqvist S, Aslett M, Dougan G, et al. Identification and characterization of the novel colonization factor CS30 based on whole genome sequencing in enterotoxigenic Escherichia coli (ETEC). Scientific reports. 2017;7:465.

9. Gaastra W, Svennerholm A-M. Colonization factors of human enterotoxigenic Escherichia coli (ETEC). Trends Microbiol. 1996;4:444–52.

10. 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

695 Escherichia coli adhesive fimbriae. Journal of clinical microbiology. 2011;49:1403–
696 10.

697 11. Cádiz L, Torres A, Valdés R, Vera G, Gutiérrez D, Levine MM, et al. Coli
698 Surface Antigen 26 Acts as an Adherence Determinant of Enterotoxigenic
699 Escherichia coli and Is Cross-Recognized by Anti-CS20 Antibodies. Frontiers in
700 microbiology. 2018;9:248.

701 12. Canto FD, O'Ryan M, Pardo M, Torres A, Gutiérrez D, Cádiz L, et al. Chaperone-
702 Usher Pili Loci of Colonization Factor-Negative Human Enterotoxigenic Escherichia
703 coli. Frontiers in Cellular and Infection Microbiology. 2017;6:CD009029.

704 13. Grewal HM, Valvatne H, Bhan MK, Dijk L van, Gaastra W, Sommerfelt H. A
705 new putative fimbrial colonization factor, CS19, of human enterotoxigenic
706 Escherichia coli. Infect Immun. 1997;65:507–13.

707 14. Pichel M, Binsztein N, Viboud G. CS22, a novel human enterotoxigenic
708 Escherichia coli adhesin, is related to CS15. Infection and immunity. 2000;68:3280–
709 5.

710 15. Joffre E, von Mentzer A, Ghany MAE, Oezguen N, Savidge T, Dougan G, et al.
711 Allele variants of enterotoxigenic Escherichia coli heat-labile toxin are globally
712 transmitted and associated with colonization factors. Christie PJ, editor. Journal of
713 bacteriology. 2015;197:392–403.

714 16. Bolin I, Wiklund G, Qadri F, Torres O, Bourgeois AL, Savarino S, et al.
715 Enterotoxigenic Escherichia coli with STh and STp genotypes is associated with
716 diarrhea both in children in areas of endemicity and in travelers. Journal of clinical
717 microbiology. 2006;44:3872–7.

718 17. Joffre E, von Mentzer A, Svennerholm A-M, Sjöling A. Identification of new
719 heat-stable (STa) enterotoxin allele variants produced by human enterotoxigenic
720 Escherichia coli (ETEC). International journal of medical microbiology : IJMM.
721 2016;306:586–94.

722 18. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, et
723 al. Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term
724 global distribution. Nature genetics. 2014;46:1321–6.

725 19. Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M, Cunningham
726 AF, et al. A commensal gone bad: complete genome sequence of the prototypical
727 enterotoxigenic Escherichia coli strain H10407. Journal of bacteriology.
728 2010;192:5822–31.

729 20. Smith P, Lindsey RL, Rowe LA, Batra D, Stripling D, Garcia-Toledo L, et al.
730 High-Quality Whole-Genome Sequences for 21 Enterotoxigenic Escherichia coli
731 Strains Generated with PacBio Sequencing. Genome announcements. 2018;6:6167.

21. 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;190:6881–93.

22. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2018;24:335–41.

23. Quainoo S, Coolen JPM, Hijum SAFT van, Huynen MA, Melchers WJG, Schaik W van, et al. Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. Clinical microbiology reviews. 2017;30:1015–63.

24. 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. Payne SM, editor. Infection and immunity. 2011;79:950–60.

25. Sahl JW, Rasko DA. Analysis of global transcriptional profiles of enterotoxigenic Escherichia coli isolate E24377A. Infection and immunity. 2012;80:1232–42.

26. 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.

27. Sahl JW, Sistrunk JR, Baby NI, Begum Y, Luo Q, Sheikh A, et al. Insights into enterotoxigenic Escherichia coli diversity in Bangladesh utilizing genomic epidemiology. Scientific reports. 2017;7:3402.

28. Begum YA, Rydberg HA, Thorell K, Kwak Y-K, Sun L, Joffre E, et al. In SituAnalyses Directly in Diarrheal Stool Reveal Large Variations in Bacterial Load and Active Toxin Expression of Enterotoxigenic Escherichiacoliand Vibrio cholerae. Limbago BM, editor. mSphere. 2018;3:e00517-17.

29. Li B, Sun J, Han L, Huang X, Fu Q, Ni Y. Phylogenetic groups and pathogenicity island markers in fecal Escherichia coli isolates from asymptomatic humans in China. Applied and environmental microbiology. 2010;76:6698–700.

30. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nature reviews microbiology. 2010;8:207–17.

31. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the Escherichia coli phylogenetic group. Applied and environmental microbiology. 2000;66:4555–8.

32. Beghain J, Bridier-Nahmias A, Nagard HL, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microbial genomics. 2018;4:690.

769 33. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The
770 comprehensive antibiotic resistance database. Antimicrobial agents and
771 chemotherapy. 2013;57:3348–57.

772 34. Carattoli A, Zankari E, García-Fernández A, Larsen MV, Lund O, Villa L, et al.
773 PlasmidFinder and pMLST: in silico detection and typing of plasmids. Antimicrobial
774 agents and chemotherapy. 2014;58:AAC.02412-14-3903.

775 35. Kumar P, Luo Q, Vickers TJ, Sheikh A, Lewis WG, Fleckenstein JM. EatA, an
776 Immunogenic Protective Antigen of Enterotoxigenic Escherichia coli, Degrades
777 Intestinal Mucin. Payne SM, editor. Infection and immunity. 2014;82:500–8.

778 36. Patel SK, Dotson J, Allen KP, Fleckenstein JM. Identification and molecular
779 characterization of EatA, an autotransporter protein of enterotoxigenic Escherichia
780 coli. Infection and immunity. 2004;72:1786–94.

781 37. Fleckenstein JM, Roy K, Fischer JF, Burkitt M. Identification of a two-partner
782 secretion locus of enterotoxigenic Escherichia coli. Infection and immunity.
783 2006;74:2245–58.

784 38. Roy K, Hilliard GM, Hamilton DJ, Luo J, Ostmann MM, Fleckenstein JM.
785 Enterotoxigenic Escherichia coli EtpA mediates adhesion between flagella and host
786 cells. Nature. 2009;457:594–8.

787 39. Hibberd ML, McConnell MM, Willshaw GA, Smith HR, Rowe B. Positive
788 regulation of colonization factor antigen I (CFA/I) production by enterotoxigenic
789 Escherichia coli producing the colonization factors CS5, CS6, CS7, CS17, PCFO9,
790 PCFO159:H4 and PCFO166. Journal of general microbiology. 1991;137:1963–70.

791 40. Pilonieta MC, Bodero MD, Munson GP. CfaD-dependent expression of a novel
792 extracytoplasmic protein from enterotoxigenic Escherichia coli. Journal of
793 bacteriology. 2007;189:5060–7.

794 41. Joffre E, Nicklasson M, Álvarez-Carretero S, Xiao X, Sun L, Nookaew I, et al.
795 The bile salt glycocholate induces global changes in gene and protein expression and
796 activates virulence in enterotoxigenic Escherichia coli. Scientific reports. 2019;9:108.

797 42. Nicklasson M, Sjöling Å, von Mentzer A, Qadri F, Svennerholm A-M. Expression
798 of colonization factor CS5 of enterotoxigenic Escherichia coli (ETEC) is enhanced in
799 vivo and by the bile component Na glycocholate hydrate. Hensel M, editor. PLoS
800 One. 2012;7:e35827.

801 43. Sjöling Å, Wiklund G, Savarino SJ, Cohen DI, Svennerholm A-M. Comparative
802 analyses of phenotypic and genotypic methods for detection of enterotoxigenic
803 Escherichia coli toxins and colonization factors. Journal of clinical microbiology.
804 2007;45:3295–301.

805 44. Wachsmuth K, Wells J, Shipley P, Ryder R. Heat-labile enterotoxin production in
806 isolates from a shipboard outbreak of human diarrheal illness. Infect Immun.
807 1979;24:793–7.

808 45. Hazen TH, Nagaraj S, Sen S, Permala-Booth J, Canto FD, Vidal R, et al. Genome
809 and Functional Characterization of Colonization Factor Antigen I- and CS6-Encoding
810 Heat-Stable Enterotoxin-Only Enterotoxigenic Escherichia coli Reveals Lineage and
811 Geographic Variation. Overall CM, editor. mSystems. 2019;4:209.

812 46. Rozwandowicz M, Brouwer MSM, Fischer J, Wagenaar JA, Gonzalez-Zorn B,
813 Guerra B, et al. Plasmids carrying antimicrobial resistance genes in
814 Enterobacteriaceae. The Journal of antimicrobial chemotherapy. 2018;73:1121–37.

815 47. EUCAST EC for AST. Determination of minimum inhibitory concentrations
816 (MICs) of antibacterial agents by broth dilution. John Wiley & Sons, Ltd (10.1111);
817 2003 Aug p. ix–xv.

818 48. Waters SH, Rogowsky P, Grinsted J, Altenbuchner J, Schmitt R. The tetracycline
819 resistance determinants of RP1 and Tn1721: nucleotide sequence analysis. Nucleic
820 acids research. 1983;11:6089–105.

821 49. Octavia S, Sara J, Lan R. Characterization of a large novel phage-like plasmid in
822 Salmonella enterica serovar Typhimurium. FEMS Microbiology Letters. 2015;

823 50. Liu P, Li P, Jiang X, Bi D, Xie Y, Tai C, et al. Complete genome sequence of
824 Klebsiella pneumoniae subsp. pneumoniae HS11286, a multidrug-resistant strain
825 isolated from human sputum. Journal of bacteriology. 2012;194:1841–2.

826 51. Falgenhauer L, Yao Y, Fritzenwanker M, Schmiedel J, Imirzalioglu C,
827 Chakraborty T. Complete Genome Sequence of Phage-Like Plasmid pECOH89,
828 Encoding CTX-M-15. Genome announcements. 2014;2:2227.

829 52. Colavecchio A, Jeukens J, Freschi L, Rheault J-GE, Kukavica-Ibrulj I, Levesque
830 RC, et al. Complete Genome Sequences of Two Phage-Like Plasmids Carrying the
831 CTX-M-15 Extended-Spectrum β-Lactamase Gene. Genome announcements.
832 2017;5:90.

833 53. Jacoby GA. Mechanisms of Resistance to Quinolones. Clin Infect Dis.
834 2005;41:S120–6.

835 54. Gilcrease EB, Casjens SR. The genome sequence of Escherichia coli tailed phage
836 D6 and the diversity of Enterobacteriales circular plasmid prophages. Virology.
837 2018;515:203–14.

838 55. Beatty ME, Bopp CA, Wells JG, Greene KD, Puhr ND, Mintz ED. Enterotoxin-
839 producing Escherichia coliO169:H41, United States. Emerging infectious diseases.
840 2004;10:518–21.

841 56. Cho S, Gupta SK, McMillan EA, Sharma P, Ramadan H, Jové T, et al. Genomic
842 Analysis of Multidrug-Resistant Escherichia colifrom Surface Water in Northeast
843 Georgia, United States: Presence of an ST131 Epidemic Strain Containing blaCTX-
844 M-15on a Phage-Like Plasmid. Microbial Drug Resistance. 2019;mdr.2019.0306.

845 57. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, et al. Cryptic
846 prophages help bacteria cope with adverse environments. Nat Commun. 2010;1:147.

847 58. Hu WS, Lin J-F, Lin Y-H, Chang H-Y. Outer membrane protein STM3031
848 (Ail/OmpX-like protein) plays a key role in the ceftriaxone resistance of Salmonella
849 enterica serovar Typhimurium. Antimicrob Agents Ch. 2009;53:3248–55.

850 59. Gómez-Duarte OG, Kaper JB. A plasmid-encoded regulatory region activates
851 chromosomal eaeA expression in enteropathogenic Escherichia coli. Infect Immun.
852 1995;63:1767–76.

853 60. Murray EL, Conway T. Multiple Regulators Control Expression of the Entner-
854 Doudoroff Aldolase (Eda) of Escherichia coli. J Bacteriol. 2005;187:991–1000.

855 61. Villa L, García-Fernández A, Fortini D, Carattoli A. Replicon sequence typing of
856 IncF plasmids carrying virulence and resistance determinants. The Journal of
857 antimicrobial chemotherapy. 2010;65:2518–29.

858 62. Osborn AM, Tatley FM da S, Steyn LM, Pickup RW, Saunders JR. Mosaic
859 plasmids and mosaic replicons: evolutionary lessons from the analysis of genetic
860 diversity in IncFII-related replicons. Microbiology+. 2000;146:2267–75.

861 63. Nishikawa Y, Helander A, OGASAWARA J, MOYER NP, HANAOKA M,
862 HASE A, et al. Epidemiology and properties of heat-stable enterotoxin-producing
863 Escherichia coliserotype O169[ratio]H41. Epidemiology and infection. 1998;121:31–
864 42.

865 64. Torres OR, González W, Lemus O, Pratdesaba RA, Matute JA, Wiklund G, et al.
866 Toxins and virulence factors of enterotoxigenic Escherichia coli associated with
867 strains isolated from indigenous children and international visitors to a rural
868 community in Guatemala. Epidemiol Infect. 2014;143:1662–71.

869 65. Sack DA, Shimko J, Torres O, Bourgeois AL, Francia DS, Gustafsson B, et al.
870 Randomised, double-blind, safety and efficacy of a killed oral vaccine for
871 enterotoxigenic E. Coli diarrhoea of travellers to Guatemala and Mexico. Vaccine.
872 2007;25:4392–400.

873 66. Ban E, Yoshida Y, Wakushima M, Wajima T, Hamabata T, Ichikawa N, et al.
874 Characterization of unstable pEntYN10 from enterotoxigenic Escherichia coli
875 (ETEC) O169:H41. Virulence. 2015;6:735–44.

876 67. Pfeifer E, Sousa JAM de, Touchon M, Rocha EPC. Bacteria have numerous
877 phage-plasmid families with conserved phage and variable plasmid gene repertoires.
878 Biorxiv. 2020;2020.11.09.375378.

879 68. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al.
880 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing
881 data. Nature methods. 2013;10:563–9.

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

885 70. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
886 assemblies from short and long sequencing reads. Phillippy AM, editor. PLoS
887 computational biology. 2017;13:e1005595.

888 71. Page AJ, Ainsworth EV, Langridge GC. socru: typing of genome-level order and
889 orientation around ribosomal operons in bacteria. Microb Genom. 2020;

890 72. Patel IR, Gangiredla J, Mammel MK, Lampel KA, Elkins CA, Lacher DW. Draft
891 Genome Sequences of the Escherichia coli Reference (ECOR) Collection. Microbiol
892 Resour Announc. 2018;7:e01133-18.

893 73. Horesh G, Blackwell GA, Tonkin-Hill G, Corander J, Heinz E, Thomson NR. A
894 comprehensive and high-quality collection of Escherichia coli genomes and their
895 genes. Microb Genom. 2021;

896 74. Kuhlmann FM, Martin J, Hazen TH, Vickers TJ, Pashos M, Okhuysen PC, et al.
897 Conservation and global distribution of non-canonical antigens in Enterotoxigenic
898 Escherichia coli. Torres AG, editor. PLoS neglected tropical diseases.
899 2019;13:e0007825.

900 75. Rasko DA, Canto FD, Luo Q, Fleckenstein JM, Vidal R, Hazen TH. Comparative
901 genomic analysis and molecular examination of the diversity of enterotoxigenic
902 Escherichia coli isolates from Chile. Plos Neglect Trop D. 2019;13:e0007828.

903 76. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al.
904 ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads.
905 bioRxiv. 2017;118000.

906 77. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary:
907 Rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.

908 78. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites:
909 rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial
910 genomics. 2016;2:e000056.

911 79. Nguyen L-T, Schmidt HA, Haeseler A von, Minh BQ. IQ-TREE: A Fast and
912 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
913 Mol Biol Evol. 2015;32:268–74.

914 80. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc
915 Bioinform. 2020;69:e96.

916 81. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.
917 2014;30:2068–9.

82. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinform Oxf Engl. 2011;28:464–9.

83. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. Bmc Bioinformatics. 2009;10:421.

84. Alikhan N-F, Petty NK, Zakour NLB, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. Bmc Genomics. 2011;12:402.

85. Li X, Xie Y, Liu M, Tai C, Sun J, acids ZDN, et al. oriTfinder: a web-based tool for the identification of origin of transfers in DNA sequences of bacterial mobile genetic elements. academic.oup.com.

86. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic acids research. 2017;45:D566–73.

87. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. Bioinform Oxf Engl. 2017;34:292–3.

88. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67:2640–4.

89. Seemann T. Abricate [Internet]. Available from: https://github.com/tseemann/abricate

90. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids research. 2016;44:W16-21.

91. Partridge SR, Tsafnat G. Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database. J Antimicrob Chemother. 2018;73:883–90.

92. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34:D32–6.

952

953

954

955

956

Acknowledgements

No acknowledgements to mention.

Authors' contributions

AvM conceived and designed the experiments, performed the experiments, analysed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored the paper and approved the final draft.

GB and AvM annotated all IS elements, transposons as well as other mobile elements, contributed to the paper and approved the final draft.

DP and AvM annotated the identified prophages, contributed to the paper and approved the final draft.

CB performed the *in silico* analysis of the genomic antibiotic resistance profiling, contributed to the paper and approved the final draft.

EJ performed the antibiotic resistance profiling, contributed to the paper and approved the final draft.

AJP assembled the genomes, contributed to the paper and approved the final draft.

AMS conceived and designed the experiments contributed to the paper and approved the final draft.

GD conceived and designed the experiments and approved the final draft.

ÅS conceived and designed the experiments, analysed data, authored the paper and approved the final draft.

Competing interests

The authors declare that there are no conflicts of interest.

Data availability

The datasets supporting the conclusions of this article are included within the articles and its additional files. The sequencing data generated in this study has been submitted to EMBL (Additional file 4 and 5). The physical ETEC reference strains can be requested by contacting the corresponding author Astrid von Mentzer (avm@sanger.ac.uk or mentzerv@chalmers.se). The database used for annotating ETEC virulence factors, ETEC virulence database, including the LT and ST alleles can be found in the github repositories: https://github.com/avonm/ETEC_vir_db and https://github.com/avonm/ETEC_toxin_variants_db.

An interactive version of the core genome phylogeny of the 1,065 *E. coli* and ETEC isolates along with the ETEC reference strains (Figure S10) reported here is accessible at https://microreact.org/project/2ZZzaHzeXbMEw9U2MAk7pK?tt=cr

Availability of material

Requests for obtaining clinical isolates collected as part of this study should be addressed to the corresponding author. Exchange of clinical isolates should always be in agreement with the University of Gothenburg.

Ethics declarations

Not applicable.

Consent for publication

Not applicable.

Additional information

Additional file 1: Supplemental Figures and Tables

Additional file 2: Detailed description of ETEC reference plasmids

Additional file 3: Excel file with plasmid classification – Inc groups

Additional file 4: Excel file with metadata of ETEC and *E. coli* genomes included in the phylogenetic tree

Additional file 5: Excel file with a compiled list of all accession numbers for chromosomes and plasmids

Funding

AvM, AMS and ÅS were supported by the Swedish Foundation for Strategic Research (grant nr. SB12-0072). AvM was also supported by The Swedish Research Council (grant nr. 2018-06828) and the Swedish Society for Medical Research (P18-0140). AJP was supported by the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Programme Microbes in the Food Chain BB/R012504/1. GD was supported by the Wellcome Trust (grant WT 098051).

Author details

[1]Department of Microbiology and Immunology, Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 7, 413 90 Gothenburg, Sweden

[2]Wellcome Sanger Institute, Hinxton, CB10 1SA, Cambridge, UK

[3]EMBL-EBI, Hinxton, CB10 1SA, Cambridge, UK

[4] University of Cambridge Department of Medicine, Addenbrookes Hospital Box 157, Hills Rd, CB2 0QQ, Cambridge, UK

[5] Department of Microbiology, Tumor and Cell Biology, Biomedicum, Karolinska Institutet, Solnavägen 9, 171 65 Solna, Stockholm, Sweden

[6]Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ Norwich, UK

[#]Currently affiliated with Chalmers University of Technology, Chalmers Tvärgata 3, 413 58 Gothenburg, Sweden

Figures

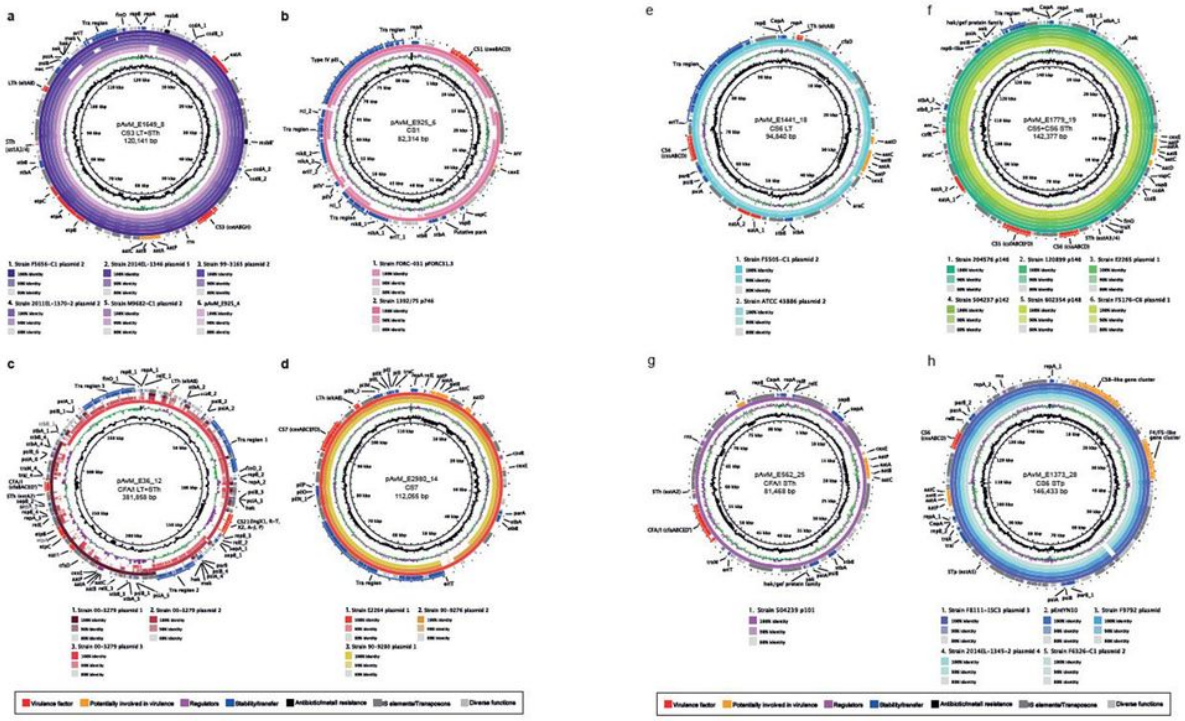


Figure 1

Comparison between the ETEC reference plasmids harbouring colonisation factors and other PacBio-sequenced ETEC plasmids using blastn. a) pAvM_E1649_8 (CS3) as reference and pAvM_925_4 (CS3) compared to the following ETEC plasmids: F5656-C1 plasmid 2 (USA, CP024262.1), 2014-EL-1346-6 plasmid 5 (2014, USA; CP024237.1), 99-3165 plasmid 2 (USA; CP029980.1), 2011EL-1370-2 plasmid 2 (2011, USA; CP022914.1) and M9682-C1 plasmid 2 (1975, USA; CP024277.1). b) pAvM_E925_6 (CS1) compared to ETEC plasmids pFORC31.3 (2004, Korea; CP013193.1) and 1392/75 p746 (1973; FN822748.1). c) pAvM_E36_12 (CFA/I) compared to plasmids 1-3 (p1: CP024294.1; p2 CP024295.1; p3: CP024296.1) from ETEC strain 00-3279 (USA). d) pAvM_E2980_14 (CS7) compared to E2264 plasmid 1 (2006, Bangladesh; CP023350.1), 90-9276 plasmid 2 (1988, Bangladesh; CP024298.1) and 90-9280 plasmid 1 (1988, Bangladesh; CP024241.1). e) pAvM_E1441_18 (CS6) compared to F5505-C1 plasmid 2 (2013, Sweden; CP023259.1) and ATCC 43886 plasmid 2 (CP024255.1). f) pAvM_E1779_19 (CS5+CS6) compared to 204576 p146 (2010, Mali; CP025908.1), 120899 p146 (2012, Gambia; CP025917.1), E2265 plasmid 1 (2006, Bangladesh; CP023347.1), 504237 p142 (2010, India; CP025863.1), 602354 p148 (2009, Bangladesh; CP025848.1) and F5176-C6 plasmid 1 (1997; CP024668.1). g) pAvM_E562_25 (CFA/I) compared to p504239_101 (2010, India; CP025860.1). h) pAvM_E1373_28 (CS6) compared to F8111-1SC3 plasmid 3 (USA; CP024272.1), pEntYN10 (1991, Japan; AP014654.2), F9792 plasmid (USA; CP023274.1), 2014EL-1345-2 plasmid 4 (2014, USA; CP024227.1) and F6326-C1 plasmid 2 (1998, USA; CP024265.1). The thresholds chosen for the blastn are shown in the key below each plasmid comparison. The colour code for the annotations are listed at the bottom of the figure. The two most

inner rings depict GC content in black and GC Skew- in purple and GC Skew+ in green. The figures were generated using BRIG [84] v0.95.

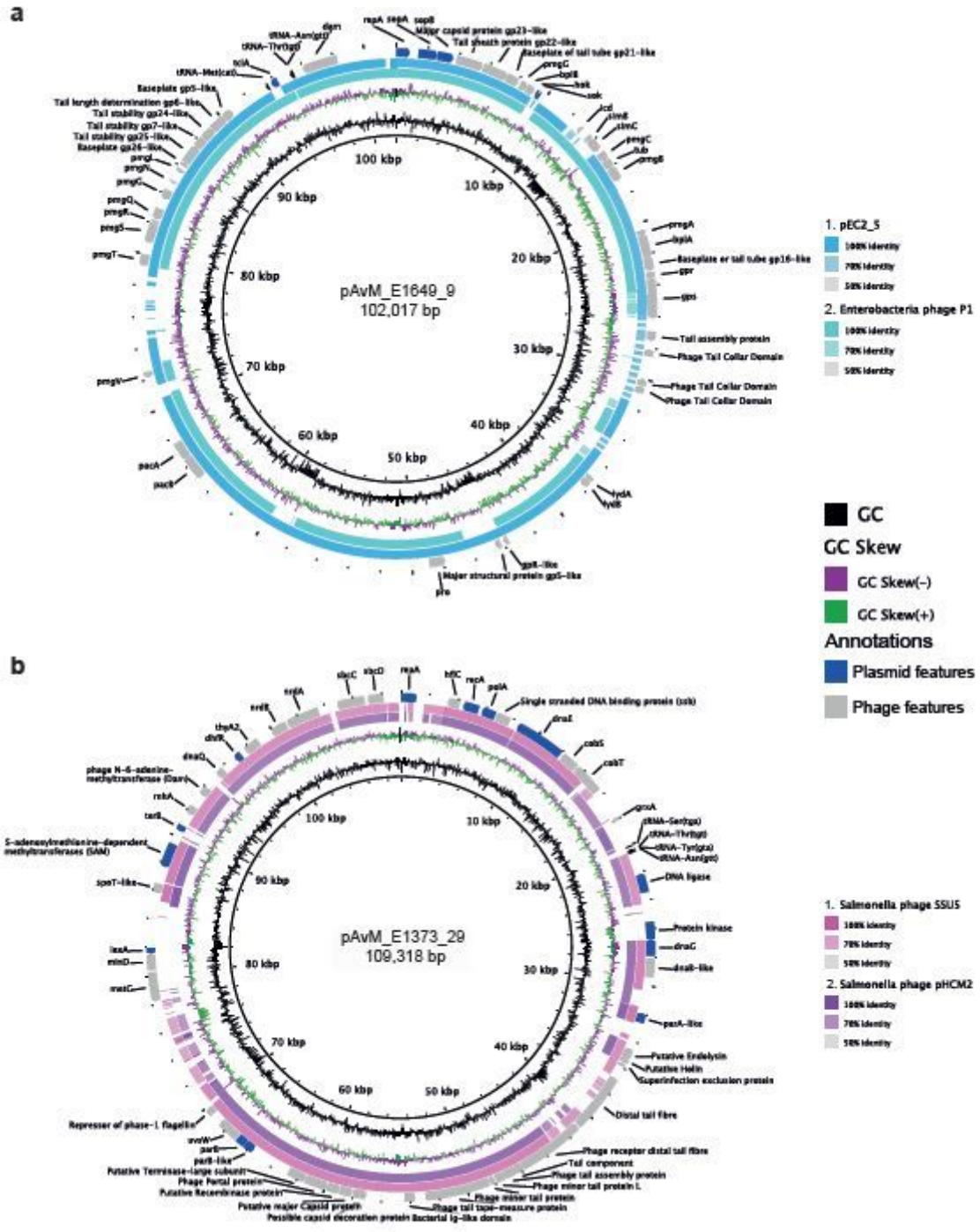


Figure 2

Comparisons between the identified ETEC phage-plasmids 361 and other similar phage-plasmids using blastn. a) pAvM_E1649_9 is a P1-like phage-plasmid here compared to Enterobacteria phage P1 (Escherichia virus P1; NC_005856.1) and pEC2_5 (E. coli strain EC2_5; CP041960.1). b) pAvM_E2980_29,

a phage-plasmid similar to the pHCM2 (Salmonella typhi strain CT18; AL513384.1) and SSU5 (Salmonella phage; JQ965645.1) in Salmonella typhi. Blastn comparisons were made using BRIG [84] (v0.95) with the thresholds indicated to the right of each plasmid comparison. Selected phage and plasmid annotations are shown in the outer ring.

Supplementary Files

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

- Table1.pdf
- Table2.pdf
- Additionalfile1.pdf
- AdditionalFile2finalPlasmiddescription.pdf
- Additionalfile3Incprofiles.xlsx
- Additionalfile4metadatav2final.xlsx
- Additionalfile5accno.xlsx