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[2, 3], Derek Pickard[4], Christine J. Boinett[2], Enrique Joffré[5], Andrew J Page[2,6], Ann-Mari Svennerholm[1], Gordon Dougan[4] and Åsa Sjöling[5]

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

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

[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

Background: 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.

Results: 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. The plasmids carrying ETEC virulence genes were compared to other available long-read sequenced ETEC strains using blastn. The ETEC reference strains harbour between two and five plasmids, including virulence, antibiotic resistance and phage-plasmids. The virulence plasmids carrying the colonisation factors are highly conserved as shown by comparison with plasmids with other ETEC strains and confirm that the plasmids and chromosomes of ETEC are both crucial for ETEC virulence and success as pathogens.

Conclusion: We confirm that the major ETEC lineages all harbour conserved plasmids that have been associated with their respective background genomes for decades. 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

Background

Diarrheal pathogens are a leading cause of morbidity and mortality globally (WHO, 2017), with enterotoxigenic *Escherichia coli* (ETEC) accounting for a large proportion

51 of the diarrhoea cases in resource-poor countries [1]. An estimation of 220 million cases
52 each year are attributed to ETEC (WHO PPC 2020). The most vulnerable group is
53 children under five years, but ETEC can also cause disease in adults and is the principal
54 cause of diarrhoea in travellers. Resource-poor settings, where access to clean water is
55 limited, enable the spread of ETEC, transmitted via the faecal-oral route through
56 ingestion of contaminated food or water [2]. The disease severity may range from mild
57 to cholera-like symptoms with profuse watery diarrhoea. The infection is usually self-
58 limiting, lasting three to four days and may be treated by water and electrolyte
59 rehydration to balance the loss of fluids and ions. There is strong evidence to support
60 that an ETEC vaccine is of key importance to prevent children and adults from
61 developing ETEC disease [3]. Several efforts are on-going to develop an ETEC
62 vaccine, with the majority focusing on including immunogenic antigens possibly
63 capable of inducing protection against a majority of the circulating ETEC clones [4–6].
64
65 ETEC bacteria adhere to the small intestine through fimbrial, fibrillar or afimbrial outer
66 membrane-structures called colonisation factors (CF). Upon colonisation, the bacteria
67 proliferate and secrete heat-labile toxin (LT) and/or heat-stable toxins, (STh or STp)
68 causing diarrhoea and often vomiting causing the further spread of the bacteria in the
69 environment [7].
70
71 The ability of an ETEC strain to infect relies on its ability to adhere to cells of a specific
72 host. To this day, 27 different CFs with human tropism have been described, and
73 individual ETEC strains usually express 1-3 different CFs [8–14]. The enterotoxins, LT
74 and ST, can also be subdivided based on structure and function. Human-associated
75 ETEC strains express one of the 28 different LT-I variants (LTh-1 and LTh-2 are the

76 most common variants) [15] alone or together with one of the genetically distinct types
77 of STa; STh and STp [16,17].
78
79 We have previously shown that ETEC strains causing human disease can be grouped
80 into a set of clonal lineages that encompass strains with specific virulence profiles.
81 Seven of the 21 identified lineages encompass ETEC strains that express the most
82 commonly found CFs and toxin profiles amongst isolated clinical ETEC strains [4,18].
83
84 There is currently one complete ETEC reference genome, H10407 [19], with curated
85 annotations. Several additional complete ETEC genomes are available [20,21], some
86 of which are annotated using automated annotation pipelines that often fail at correctly
87 annotating ETEC specific genes such as CFs. The rapid adaptation of next-generation
88 sequencing in public health, specifically within bacterial diseases [22,23] and several
89 large-scale sequencing studies [18,24–28] has led to a sharp increase in the number of
90 publicly available ETEC genomes. Most of these data were generated with short-read
91 technologies, such as Illumina. A limitation of short-read sequence data is the inability
92 to unambiguously resolve repetitive regions of a genome, leading to fragmented de
93 novo assemblies of the underlying genome, missing regions and genes, and disjointed
94 synteny. ETEC is a highly diverse pathogen both in the core genome (SNPs) and the
95 accessory genome, including mobile genetic elements (MGE). Clinically related
96 MGEs, such as virulence plasmids, vary within ETEC strains. Hence, it is important to
97 identify lineage-specific reference genomes that are carefully annotated, i.e. manually
98 curated annotations, and include both chromosome and plasmid(s). Several complete
99 genomes have been generated using long-read sequencing alone [20,28], however,
100 circularising some chromosomes and plasmids may be difficult, and small plasmids can

101 be lost. Assembly issues can be resolved using a hybrid assembly approach combining
102 long-read and short-read sequencing data. In this report, we describe eight genomes,
103 eight chromosomes (seven successfully circularised) and 29 plasmids (24 successfully
104 circularised) with curated annotations, from isolates representing the major ETEC
105 lineages (L1-L7) that cause disease globally. They are sequenced using both short and
106 long-read sequencing technologies to provide the highest accuracy currently available.
107 These reference genomes will form the foundation of ETEC genomics research for
108 years to come.
109

110 **Results**

111 ***Genome analysis of eight representative ETEC isolates***

112 Eight ETEC strains representing the seven major ETEC lineages (L1-L7) comprising
113 isolates with the most prevalent virulence factor profiles were sequenced, assembled,
114 circularised and manually curated (Table 1).
115 L3 includes two different representative strains, one CS7 and one CFA/I positive strain.
116 All chromosomes except one were circularised (E1779). The average length of the
117 chromosome was 4,927,521 bases (4,721,269-5,151,162) with an average GC content
118 of 50.7% (50.4% to 50.9%) and the number of CDS ranging from 4,409 to 4,924 (Table
119 S1). Each ETEC reference genome contains between two and five plasmids
120 encompassing plasmid-specific features. Some of which carried virulence genes and/or
121 antibiotic resistance genes (Table 2, Additional File 2).
122

123 ***Comparative genomics of the chromosome***

124 The chromosomes of the reference strains were aligned and compared using MAUVE,
125 and the overall structure is conserved across all eight chromosomes (Figure S1). In

126 total, 8,348 chromosomal genes were identified in the eight ETEC strains with 3,179
127 genes considered part of the core genome shared by all eight reference strains. The
128 majority of human commensal *Escherichia coli* (*E. coli*) strains belong to subgroup A
129 [29,30]. However, ETEC strains fall into multiple phylogenetic groups (A, B1, B2, D,
130 E, F and CladeI with the majority found in the phylogenetic groups A and B1 [18]. The
131 phylogenetic group of the eight ETEC reference strains have previously been
132 determined using the triplex-PCR scheme [31]. The ETEC references were re-analysed
133 using ClermonTyping [32] and it was determined that strain E1373 belongs to the
134 phylogenetic group E while the other reference isolates belong to groups A and B1
135 (Table 1).
136
137 ***Plasmids***
138 The plasmids of each isolate were annotated using Prokka followed by manual curation
139 of the annotations including genes part of the conjugation machinery and known
140 plasmid stability genes. Virulence factors (including CFs, toxins and putative virulence
141 factors), antibiotic resistance determinants with the Comprehensive Antibiotic
142 Resistance Database (CARD) [33] as well as complete and partial insertion elements
143 and prophages were manually annotated. The plasmids were designated
144 pAvM_strainID_integer, e.g. pAvM_E925_4 (Additional file 2). The first plasmid
145 reported in this study starts at 4 as three previous plasmids E873p1-3 already have been
146 deposited to GenBank related to a different project [8].
147
148 Plasmids were typed by analysing the presence and variation of specific replication
149 genes to assign the plasmids to incompatibility (Inc) groups. The Inc groups of the
150 ETEC reference plasmids were first determined using PlasmidFinder and further

151 classified into subtypes using pMLST [34]. The replicons identified are IncFII,
152 IncFIIA, IncFIIS, IncFIB, IncFIC, IncI1 and IncY. Plasmids with replicon IncY,
153 IncFIIY or IncB/O/K/Z mainly harboured plasmid associated genes, such as stability
154 and transfer genes. Importantly, replicons FII, FIB and I1 were found to be strongly
155 associated with virulence genes as all CF loci, toxins and virulence factors *eatA* and
156 *etpBAC* were present on these plasmids. The majority of all ETEC plasmids analysed
157 here (17/29) belong to IncFII, of which six of the IncFII plasmids have an additional
158 IncFIB replicon. In six of the ETEC reference strains two or three IncFII replicons are
159 present, for example, in strain E925, the plasmids pAvM_E925_4 and 7 both belong to
160 IncFII. However, the plasmids were further subtyped to FII-111 and FII-15,
161 respectively, (Table 2 and Additional file 3), explaining the plasmid compatibility.
162
163 ***Virulence factors***
164 The CFs expressed by the selected reference strains are CFA/I, CS1-CS3, CS5-CS7 and
165 CS21. Three of the strains (E925, E1649 and E1779) express both LT and ST, two
166 strains (E2980 and E1441) express LT and the strains E36 and E562 express STh, while
167 E1373 express STp (Table 1). A plasmid can harbour multiple virulence genes, usually
168 a CF locus and genes encoding one or two toxins. Interestingly, plasmids do not often
169 harbour multiple CF loci, but on individual plasmids (in the ETEC reference strains
170 described here). Exceptions for this is strain E1779 in which CS5 and CS6 loci are
171 located on the same plasmid (pAvM_E1779_19). In both E925 (L1) and E1649 (L2)
172 the genes encoding CS3 (*cstABGH*), ST (*estA*) and LT (*eltAB)* are located on the same
173 plasmid, both with the FII replicon and of roughly the same size (Table 2). Blastn
174 comparison between the plasmids and additional plasmids that harbour the same
175 virulence genes shows that they are highly conserved (Figure 1). The results correspond

176 with the close genetic relationship and common ancestry of lineage 1 (L1) and lineage
177 2 (L2) [18].
178 Besides CFs and toxins, additional virulence factors were identified in the majority of
179 the strains (Table 2), with *eatA* and *etpBAC* being the most commonly found.
180 EatA is an immunogenic mucinase that contributes to virulence by degrading MUC2
181 which is the major protein component of mucus in the small intestine [35,36]. The
182 *etpABC* genes encode an adhesin located on the tip of the flagella and mediate
183 adherence to host cells [37,38]. Four reference strains (E925, E1649, E36 and E562)
184 harbour both *eatA* and *etpBAC.* In three strains the *eatA* and/or *etpBAC* are located on
185 the same plasmid with an FII or FII+FIB replicon along with additional ETEC virulence
186 genes, except in E562 and E1373, where *eatA* and *etpBA*C are located on an I1+FII
187 (pAvM_E562_23) and I1 (pAvM_E1373_16) plasmid, respectively, which mainly
188 contains plasmid associated genes including genes encoding the *pil* operon and *tra*-
189 operon (pAvM_E562_23). Furthermore, a less explored putative virulence factor is
190 CexE, which is an extracytoplasmic protein dependent on the expression of the CFA/I
191 regulator *cfaD* [39], and was first identified in H10407 [40]. Corroborating earlier
192 findings, the CFA/I positive E36 (L3) and E562 (L6) isolates harbour *cexE*
193 (pAvM_E36_12 and pAvM_E562_25). In addition, *cexE* is present in pAvM_E925_6,
194 pAvM_E1779_19 and pAVM_E2980_14, pAvM_E1441_18 and pAvM_E1373_28.
195 *CexE* has previously also been identified in several CS5+CS6 positive ETEC and
196 shown to be upregulated in the presence of bile and sodium glycocholate-hydrate [41].
197 Bile is known to be involved in the regulation of several ETEC CFs [42,43]. The
198 location of *cexE* seems to be conserved across specific strains. In pAvM_E36_12,
199 pAvM_E1441_18, pAvM_E1779_19 and pAvM_E562_25 *cexE* is located upstream of
200 the *aatPABC* locus, whereas in pAvM_E925_6 and pAvM_E2980_14 *cexE* is located

201 downstream of *rob* (an AraC family transcriptional regulator) in the opposite direction.
202 pAvM_E925_4 harbours the *aatPABC* locus; however *cexE* is located on a different
203 plasmid (pAvM_E925_6) in this strain.
204
205 ***Comparison of plasmids with the same virulence profile***
206 ETEC isolates within a lineage share the same virulence profile, specifically the same
207 CF profile (Figure S10). We verified that our selected isolates grouped within
208 previously described lineages with confirmed virulence profiles by phylogenetic
209 analyses (Figure S2). Blastn of each of the CF positive plasmids from each reference
210 genome were performed, and the best hit(s) were used for subsequent analysis (Figure
211 1). Most of the plasmids identified as related to the ETEC reference plasmids were not
212 annotated, hence, when needed these were annotated using the corresponding ETEC
213 reference plasmids annotation as a high priority when running Prokka. We show that
214 plasmids with the same CF and toxin profile from the same lineage are often conserved
215 (Figure 1). For example, the two plasmids encoding CS3 (pAvM_E925_4 and
216 pAvM_E1649_8) are highly similar to several CS3 harbouring plasmids from O6:H16
217 strains collected from various geographical locations between 1975 and 2014, including
218 *E. coli* O6:H16 strain M9682-C1 plasmid unnamed2 (CP024277.1) and *E. coli* strain
219 O6:H16 F5656C1 plasmid unnamed2 (CP024262.1) PacBio sequenced by Smith et al.
220 [20] (Figure 1a). Furthermore, high coverage and similarity were found between the
221 plasmids of isolates E1441 (L4), and PacBio sequenced plasmids of ETEC isolates
222 ATCC 43886/E2539C1 and 2014EL-1346-6 [20]. These isolates were collected in the
223 seventies [44] and 2014 (from a CDC collection), respectively, and assigned as O25:
224 H16 which is the O group determined for E1441 in silico (Figure 1e). Plasmids of
225 E2980 (LT+ CS7, L3) were validated by the PacBio sequenced plasmids of ETEC

226 isolate E2264 (Figure 1d). Similarly, two plasmids of E1779 (LT, STh+CS5+CS6, L5)
227 was identified in E2265 (LT, STh+CS5+CS6 [28,41], although E1779 harboured two
228 additional plasmids. Several additional L5 ETEC genomes have been sequenced within
229 the GEMS study [45], and high plasmid similarity and conservation in CS5+CS6
230 positive L5 isolates was evident (Figure 1f).
231
232 Overall the results show that ETEC plasmids are specific to lineages circulating
233 worldwide and conserved over time (Figure 1, Figure S2-S9 for more extensive plasmid
234 annotation and Figure S10). Thus, the plasmids of major ETEC lineages must confer
235 evolutionary advantages to their host genomes since they are seldom lost.
236
237 ***Antibiotic resistance***
238 *E. coli* can become resistant to antibiotics, both via the presence of antibiotic resistance
239 genes and the acquisition of adaptive and mutational changes in genes encoding efflux
240 pumps and porins which allows the bacterium to pump out the antibiotic molecules
241 effectively.
242
243 Antibiotic resistance genomic marker(s), both chromosomally located and on plasmids,
244 were identified using the CARD database [33] (Table 2, Figure S12 and Figure S13 and
245 Additional file 2). Similar to other studies, IncFII and B/O/K/Z plasmids were found to
246 harbour genes conferring antibiotic resistance [46]. Furthermore, the phenotypic
247 antibiotic resistance profile was determined with clinical MIC breakpoints based on
248 EUCAST (The European Committee on Antimicrobial Susceptibility Testing) [47]
249 (Table S2). Phenotypic antibiotic resistance profiles (Table S2) were supported mainly
250 by the findings of antibiotic resistance genes, efflux pumps and porins (Figures S4 and

251 S5 and Table S3), although some differences were found. All ETEC reference strains
252 are phenotypically resistant to at least two antibiotics of the 14 tested (Table S2).
253 Resistance against penicillin's, norfloxacin (Nor) and chloramphenicol (Cm) is most
254 common among these strains. Two of the strains, E1441 and E2980, harbour more than
255 four antibiotic resistance genes as well as multiple efflux systems and porins (Figure
256 S12, Figure S13 and Table S3). The plasmid pAvM_E1441_17 carries *aadA1-like*,
257 *dfrA15*, *sul1* and *tetA*(A) resistance genes (Table 2), where the first three genes are in a
258 Class 1 integron which confers resistance to streptomycin, trimethoprim, and
259 sulphonamide (sulphamethoxazole). The gene *tetA*(A) is part of a truncated Tn*1721*
260 transposon [48]. The E1441 strain was verified as resistant to tetracycline (Tet) and
261 sulphamethoxazole-trimethoprim (Sxt) while streptomycin was not tested. A *mer*
262 operon derived from Tn*21* is also present the resistance region of pAvM_E1441_17
263 (Table 2), indicating that the plasmid would also likely confer tolerance to mercury,
264 although this was not confirmed. Interestingly, this multi-replicon (FII and FIB)
265 plasmid also harbours the *lng* locus encoding CS21, one of the most prevalent ETEC
266 CFs. In isolate E2980 virulence plasmid pAvM_E2980_15 harboured multiple
267 resistance genes in the same region ($bla_{\text{TEM-1b}}$, *strA*, *strB* and *sul2)* conferring resistance
268 to ampicillin, streptomycin and sulphonamides. E2980 was found to be resistant to
269 ampicillin (Amp) and oxacillin (Oxa), which can be broken down by the beta-lactamase
270 $\text{Bla}_{\text{TEM-1b}}$, (Table 2, Table S2 and Table S3). E562 harbours three antibiotic resistance
271 genes, *ampC* located in the chromosome and the *tet*(A) and $bla_{\text{TEM-1b}}$ genes on an FII
272 plasmid (pAvM_E562_27). The *mer* operon derived from Tn*21* is also present in the
273 region (Table 2 and Table S3). The phenotypic resistance profile of E562 matches the
274 genomic profile with resistance to tetracycline (Tet), ampicillin (Amp), amoxicillin-
275 clavulanic acid (Amc) and oxacillin (Oxa) (Table S2). The plasmid pAvM_E36_13

276 contains a complete copy of Tn*10*, which encodes the *tet*(B), tetracycline resistance
277 module. Although the AvM_E1373_29 phage-like plasmid is cryptic, related plasmids
278 such as the pHMC2-family of phage-like plasmids [49] (described below), can harbour
279 resistance genes such as $bla_{\text{CTX-M-1}}$4 [50] and $bla_{\text{CTX-M-15}}$ [51,52].
280
281 Phenotypic intermediate resistance to ampicillin was found in E36 and E1779 encoded
282 by chromosomal *ampC*. Higher MIC values against ampicillin are found in E2980 and
283 E562 strains carrying bla_{TEM} genes. Phenotypic resistance to ceftazidime (Caz) and
284 ceftriaxone (Cro) was not found in the isolates, which were consistent with the absence
285 of extended-spectrum beta-lactamase (ESBL) resistance genes in the sequence data.
286
287 Resistance to chloramphenicol (Cm) was found in five isolates, but none of the resistant
288 isolates contained known resistance genes suggesting that chromosomal mutations or
289 presence of efflux pumps may account for this reduced susceptibility.
290 The ETEC reference strains contain several efflux systems which could explain why
291 the genotypic and phenotypic antibiotic resistance profile did not match for all
292 antibiotics. All of the isolates harbour multiple efflux pumps located on the
293 chromosome and plasmids (Table S3 and Figure S12). In E925, a non-synonymous
294 mutation in *acrF* was identified (G1979A) resulting in a substitution from arginine to
295 glutamine (A360Q). The effect on the expression and/or function of the AcrEF efflux
296 pump was not verified.
297
298 Phenotypic resistance to norfloxacin (Nor) was found in 6 of the isolates. The isolates
299 were analysed for chromosomal mutations likely to confer quinolone resistance, using
300 ResFinder but mutations in *gyrA* were only found in one strain, E2980, at position S83A

301 which may confer resistance to nalidixic acid, norfloxacin and ciprofloxacin. However,
302 E2980 was sensitive to Nalidixic acid. Both mutation(s) that alter the target (*gyrA* and
303 *parC*), as well as the presence of efflux pumps, can confer resistance to
304 fluoroquinolones. The majority of the isolates are moderately resistant to Norfloxacin
305 (and Nalidixic acid), both quinolones, which is most likely due to the presence of two
306 efflux pumps, AcrAB-R and AcrEF-R, as only one mutation was identified in *gyrA* of
307 isolate E2980 where usually at least two or more mutations are needed to confer
308 augmented resistance [53].
309
310 ***Identification of phage-like plasmids in ETEC***
311 Two of the ETEC reference strains (E1649 and E1373) harboured phage-like plasmids
312 (pAvM_E1649_9 and pAvM_E1373_29) which encode for DNA metabolism, DNA
313 biosynthesis as well as structural bacteriophage genes (capsid, tail etc.). Both
314 pAvM_E1649_9 and pAvM_E1373_29 contain genes associated with plasmid
315 replication, division and maintenance (i.e. *repA* and *parAB*). Phage-like plasmids are
316 found in various bacterial species, such as *E. coli*, *Klebsiella pneumoniae*, *Yersinia*
317 *pestis*, *Salmonella enterica* serovar Typhi, *Salmonella enterica* serovar Typhimurium,
318 *Salmonella enterica* serovar Derby and *Acinetobacter baumanii* [54]. The plasmid
319 pAvM_E1649_9 belong to the P1 phage-like plasmid family (Figure 2a and Figure
320 S14a) while pAvM_E1373_29 belongs to the pHCM2-family (Figure 2b and Figure
321 S14b) that can be traced back to a likely phage origin similar to the *Salmonella* phage,
322 SSU5 [49]. Both phage-plasmids thus contain replication and/or partition genes of
323 plasmid origin and a complete set of genes that are phage related in function and
324 properties (Figure 2 and Figure S14). Significantly, phage-like plasmid
325 pAvM_E1373_29 falls more within the *E. coli* lineage of pHCM2 phage-like plasmid

326 rather than those found in *Salmonella* species. This indicates that phage-like plasmids
327 have diversified within the bacterial species they were isolated.
328
329 Blastn searches confirmed high similarity (at least 80% at the DNA across much of
330 the sequence) of pAvM_E1373_29 to several phage-like plasmids found in *E. coli*
331 including ETEC O169:H41 isolate F8111-1SC3 [20,55], several $bla_{\text{CTX-M-15}}$ positive
332 phage-like plasmids (pANCO1, pANCO2 [52] and PV234a), as well as a plasmid found
333 in *E. coli* ST648 from wastewater and ST131 isolate SC367ECC [56]. The P1 phage-
334 like plasmid pAvM_1649_9 is most similar to p1107-99K, pEC2_5 isolated from
335 human urine and p2448-3 from a UPEC ST131 isolate isolated from blood. The
336 similarity is most pronounced at the amino acid level. Conservation and synteny are
337 evident when pAvM_1649_9 is compared to P1 phage.
338
339 ***Prophages present and their cargo genes***
340 Prophages may insert into chromosomes and bring along genes required for lysogeny
341 and lytic cycles and cargo genes that are often picked up when DNA is compacted into
342 the capsid. Cargo genes can significantly benefit the host bacterium by providing
343 additional elements to defence against phage or immune evasion and finally,
344 environmental survival. PHASTER analyses identified prophages in the chromosomes
345 of all ETEC reference isolates and some of the plasmids (Table S4). Putative tellurite
346 resistance operons in isolates E925, E36, E2980 and E1373 were all located in
347 prophages. In addition, *eatA* (in E925, and E1649) and *estA* (STh) genes (E36) were
348 prophage cargo genes.
349

350 Many prophage cargo genes identified in this study have properties related to inhibition
351 of cell division. Among these are a variety of *kil* genes which can enhance host bacterial
352 survival in the presence of some antibiotics [57]. Some genes that are core entities
353 within many prophages, such as *zapA* (from E1779_Pph_6), *dicB* and *dicC* (found in
354 phage E1779_Pph_7), also have similar effects as they can inhibit cell division in the
355 presence of antibiotics which raise the broader question in terms of how they are
356 beneficial to the host bacterium.
357
358 A different gene of interest is the *yfdR* gene identified in E1779_Pph_7 (gene
359 E1779_04412). YfdR curtails cellular division by inhibiting DNA replication under
360 stress conditions encountered by the bacterial cell. Similarly, the *iraM* gene located in
361 phage E1441_Pph_2 plays a role in RpoS stability.
362
363 OmpX homologs were found in numerous phages in this study. They are trans-
364 membrane located and play a role in virulence as well as antibiotic resistance [58]. PerC
365 is often associated with EPEC plasmids, where it seems to have a regulatory role for
366 the attaching and effacing gene, *eaeA* [59]. Its presence as cargo within an ETEC strain
367 phage, E1779_Pph_7 located on the chromosome, is intriguing. Its ability to regulate
368 other virulence genes is yet to be determined. Within the same phage, a *gntR*-like
369 regulatory gene was identified. This gene plays a role in gluconate utilisation and
370 induction of the Entner-Doudoroff pathway [60].
371
372 **Discussion**
373 ETEC strains have previously been shown to fall into globally spread genetically
374 conserved lineages which encompass strains with specific virulence factor profiles [18].

375 The currently widely used ETEC reference strains H10407 (CFA/I) and E24377A
376 (CS1+CS3) are highly divergent from other strains with the same virulence profile
377 sequenced more recently [18] and highlights the need for relevant and representative
378 ETEC reference strains and genomes. The long-read sequenced isolates presented here
379 comprise complete reference genomes with separate chromosomal and plasmid
380 sequences that allow more detailed studies of ETEC and *E. coli* phylogeny. The
381 reference strains are representative isolates of their respective lineage and cluster
382 phylogenetically together with different ETEC isolates sequenced by several other
383 groups (Figure S10).

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

390 Blast analyses confirm that the plasmids identified in this study are often highly
391 homologous to other plasmids identified by either long-read or Illumina sequencing
392 present in GenBank. For instance, the 94.5 kb plasmid pAvM_1441_18 was 98%
393 identical to two 96 kb and 82 kb plasmids belonging to ETEC O25: H16 isolates ATCC
394 43886/E2539C1 and 2014EL-1346-6 sequenced by PacBio by Smith *et al.,* [20] (Figure
395 1e and Figure S6). Plasmid pAvM_E1441_18 is the major virulence plasmid of this
396 lineage carrying genes encoding LT and CS6.

397 The larger plasmid in E1441 (pAvM_E1441_17) carries both the genes for ETEC CF
398 CS21 and antibiotic resistance determinants. Furthermore, complete conjugation

399 machinery was present suggesting that this is most likely a self-transmissible plasmid,
400 though this was not confirmed. Movement of such a plasmid would result in the spread
401 of ETEC virulence genes and AMR determinants.

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

419 The ETEC O169:H41 isolate F8111-1SC3 plasmid unnamed 2 [20,55] is highly similar
420 to pAvM_E1373_28 (Figure 1h and Figure S9). The F8111-ISC3 isolate is part of a
421 CDC collection of ETEC isolates from cruise ship outbreaks and diarrheal cases in US
422 1996-2003. The antibiotic resistance profiles of these isolates were determined [55] and
423 most isolates of O group 169 were tetracycline resistant consistent with the findings of

424 the *tet* gene in E1373 isolated in Indonesia in 1996. ETEC diarrhoea caused by
425 O169:H41 and STp CS6 isolates is repeatedly reported to cause diarrhoea, particularly
426 in Latin America [45,63–65]. Among the cruise ship isolates is the sequenced and
427 characterised virulence plasmid pEntYN10 encoding STp and CS6, described as
428 unstable and easily lost in vitro [63,66]. The E1373 plasmid; AvM_E1373_28 is highly
429 homologous to pEntYN10 (Figure 1h and Figure S9) and the virulence profile of ETEC
430 O169: H41 is conserved in isolates collected globally. Hence, the instability of the
431 plasmid is incongruent with current data indicating that plasmids are stable within this
432 lineage and serotype.

433 Interestingly, two distinctive extra-chromosomal elements which are highly similar to
434 P1 and SSU5 phage were identified among the 8 ETEC reference strains sequenced
435 (Figure 2, Figure S14 and Table S4). The SSU5-like element carries several genes that
436 allow it to be functional as a plasmid and belongs to the pHCM2-like family of Phage-
437 Plasmids (Figure 2b) [49]. These plasmids are devoid of virulence factors, transposons
438 and antibiotic markers but, they contain a significant number of DNA metabolism and
439 biosynthesis genes and they may contain bacteriophage inhibitory genes that have not
440 yet been identified. Interestingly, several SSU5 phage-like plasmids have been shown
441 to carry the ESBL gene $bla_{\text{CTX-M15}}$ in extra-intestinal pathogenic *E. coli* isolates [51].
442 ESBL resistance seems to be absent or low in ETEC and the SSU5 phage-like plasmid
443 pAvM_E1373_29 does not contain antibiotic resistance genes. A recent study
444 investigating the distribution of phage-plasmids show that the phage homologs tend to
445 be more conserved and the plasmid homologs more variable [67]. This is also seen in
446 the phage-plasmids identified here, e.g., genes that could be advantageous to the host
447 cell linked to metabolism and biosynthesis.
448

449 **Conclusion**

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

459

460 **Methods**

461 ***Selection of strains***

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

473 annotated, with emphasis on the plasmids. The physical ETEC reference strains are
474 available upon request.
475
476 ***Phenotypic toxin and CF analyses***
477 ETEC isolates were identified by culture on MacConkey agar followed by an analysis
478 of LT and ST toxin expression using GM1 ELISAs [43]. The expression of the different
479 CFs was confirmed by dot-blot analysis [43]. Isolates had been kept in glycerol stocks
480 at −70 °C, and each strain has been passaged as few times as possible.
481
482 ***Antibiotic susceptibility testing***
483 All ETEC isolates were tested against 14 antimicrobial agents and their minimum
484 inhibitory concentration was determined by broth microdilution using EUCAST
485 methodology [47]. The antimicrobial agents were: ampicillin, amoxicillin-clavulanic,
486 oxacillin, ceftazidime, ceftriaxone, doxycycline, tetracycline, nalidixic acid,
487 norfloxacin, azithromycin, erythromycin, chloramphenicol, nitrofurantoin and
488 sulfamethoxazole-trimethoprim. All antibiotics were purchased from Sigma-Aldrich.
489 The *E. coli* ATCC 25922 was used as quality control. The MIC was recorded visually
490 as the lowest concentration of antibiotic that completely inhibits growth.
491
492 ***DNA extraction and sequencing***
493 Strains from each lineage (L1-L7) were SMRT-sequenced on the PacBio RSII. A
494 hybrid *de novo* assembly was performed combining the reads from both the SMRT-
495 sequenced and Illumina sequenced strains.

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

498 Isolates were cultured in CFA broth overnight at 37°C followed by cell suspension in
499 TE buffer (10 mM Tris and 1 mM EDTA pH 8.0) with 25% sucrose (Sigma) followed
500 by lysis using 10 mg/ml lysozyme (in 0.25 Tris pH 8.0) (Roche). Cell membranes were
501 digested with Proteinase K (Roche) and Sarkosyl NL-30 (Sigma) in the presence of
502 EDTA. RNase A (Roche) was added to remove RNA molecules. A phenol-chloroform
503 extraction was performed using a mixture of Phenol:Chloroform:Isoamyl Alcohol
504 (25:24:1) (Sigma) in phase lock tubes (5prime). To precipitate the DNA 2.5 volumes
505 99% ethanol and 0.1 volume 3 M NaAc pH 5.2 was used followed by re-hydration in
506 10 mM Tris pH 8.0. DNA concentration was measured using NanoDrop
507 spectrophotometer (NanoDrop). On average 10 µg for PacBio sequencing. Library
508 preparation for SMRT sequencing was prepared according to the manufacturers'
509 (Pacific Biosciences) protocol. The DNA was stored in E buffer and sequenced at the
510 Wellcome Sanger Institute. Isolates were sequenced with a single SMRTcell using the
511 P6-C4 chemistry, to a target coverage of 40–60X using the PacBio RSII sequencer.

512 ***Assembly***

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

523 the short reads from the Illumina sequencing were used in combination with the long
524 reads using Unicycler [70] to generate high-quality assemblies.
525
526 Fully circularised chromosomes and plasmids were achieved for the majority of the
527 strains. Cross-validation of the assemblies was performed where two or three strains of
528 a lineage were sequenced (Figure S15). A single assembly from each lineage was
529 chosen to act as the representative reference genome, with priority given to assemblies
530 with the most complete and circularised chromosome and plasmids. In total, one
531 chromosome and 5 out of the 29 plasmids could not be circularised (independent on the
532 two strains that were sequenced initially) out of the 8 selected representative strains.
533 These are indicated in Table 2 and Table S1. Between two and five plasmids were
534 identified in the eight strains. Shorter contigs that could not be assembled properly
535 contained phage genes and are included in the genomes and annotated as prophages
536 Table S4). Socru was used to validate the assembly of the chromosome, they all have
537 biologically valid orientation and order of rRNA operons with a type GS1.0, which is
538 seen in most *E. coli* in the public domain [71].
539
540 ***Phylogenetic tree***
541 The phylogenetic relationship between the ETEC reference genomes to other ETEC
542 and *E. coli* commensals and pathotypes was investigated. The following collections
543 were included: ETEC-362 [18], ECOR [72] and the Horesh collection [73] along with
544 additional ETEC genomes from several studies [20,24,26,27,45,74,75]. The reads of
545 identified ETEC genomes from other studies were downloaded from GenBank and
546 assembled using Velvet. Long-read sequenced ETEC genomes were included in the
547 tree and were not re-assembled. The phylogroup of the ETEC strains was determined

548 using ClermonTyping [32] (v20.03). The virulence profile of the ETEC strains was
549 determined using ARIBA [76] (v2.14.16) with default settings using the custom ETEC
550 virulence database (https://github.com/avonm/ETEC_vir_db). A total of 1,066
551 genomes was included in the phylogenetic tree. The alignment of core genes (n = 2,895)
552 identified by Roary [77] (v3.12.0) was converted to a SNP-only alignment using snp-
553 sites [78]. A phylogenetic tree was produced with IQ-TREE [79] (v1.6.10) using a GTR
554 gamma model (GTR+F+I) optimised using the built-in model test and visualised using
555 the R package ggtree [80].
556
557 ***Gene prediction, annotation and comparative analysis***
558 The final assembly was annotated using Prokka [81] (v1.14.6). The annotations of all
559 plasmids generated by Prokka were manually checked using the genome viewer
560 Artemis [82] and Geneious 2018.11.1.5 (http://www.geneious.com) together with
561 blastp. Annotations of known ETEC virulence genes (colonisation factors, toxins, *eatA*
562 and *etpBAC*) were added after blast+ [83] analysis using the reference genes available
563 in the ETEC virulence database (https://github.com/avonm/ETEC_vir_db) and their
564 annotations updated accordingly. The LT and ST alleles were determined according to
565 Joffre et al., (https://github.com/avonm/ETEC_toxin_variants_db) [15,17]. Where
566 required, PFAM domains were searched using jackhammer to back up any identified
567 protein using blastp (https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer). Blastn
568 and tblastx were used for plasmid comparison, using both NCBI website or within
569 BLAST Ring Image Generator (BRIG) [84] (v0.95).
570
571 ***Incompatibility groups***

572 Due to the discrepancy in databases two approaches was used to determine the Inc
573 groups of the 25 plasmids. PlasmidFinder was used with a threshold for minimum %
574 identity at 95% and minimum coverage of 60%. The plasmids were further
575 characterised by pMLST [34], except for IncY which are a group of prophages that
576 replicate in a similar manner as autonomous plasmids (Additional File 3). IncB/O/K/Z
577 plasmids were further typed by blastn comparison to the reference B/O (M93062), K
578 (M93063) and Z (M93064) replicons.
579
580 ***oriT prediction***
581 The location of the *oriT* in the plasmids, if present, was predicted using oriTFinder [85]
582 with Blast E-value cut-off set to 0.01.
583
584 ***Genomic antibiotic resistance profiling***
585 The identification of antibiotic resistance genes, located on both the chromosome and
586 plasmid(s) as well as the presence of efflux pumps and porins known to confer
587 resistance to antibiotics. The results were obtained by running ARIBA [76] using the
588 CARD database [86] with the default settings (minimum 90% sequence identity and no
589 length cut-off). ARIBA combines a mapping/alignment and targeted local assembly
590 approach to identify AMR genes and variants efficiently and accurately from paired
591 sequencing reads. The heatmaps were visualised using Phandango [87]. The presence
592 of chromosomal mutations in *gyrA* and *parC* was determined with ResFinder (v3.2)
593 from the Center of Genomic Epidemiology [88].
594
595 ***Virulence gene prediction***

596 The ETEC assemblies from the ETEC-NCBI collection (Additional file 4) were
597 screened using abricate [89] with default settings against the ETEC virulence database
598 (https://github.com/avonm/ETEC_vir_db) for virulence gene (including *eatA* and
599 *etpBAC*) predication. A subset of the isolates in the ETEC-NCBI dataset have
600 previously been analysed for the presence of EatA where a sample with negative PCR
601 but positive western blots were included as positive [74]. Here, only isolates harbouring
602 the *eatA* and *etpBAC* genes are considered positive.
603
604 ***Prophage prediction***
605 The complete FASTA sequence of each ETEC reference genome was searched for
606 phage genes and prophages using PHASTER (phaster.ca) [90]. The identified intact
607 prophages are listed in Table S4. All prophage contained cargo genes but only
608 recognisable genes are stated, not any hypothetical. Additional questionable and not
609 intact prophages were identified but have not been included here. The prophages have
610 been given a specific identifier name and are also annotated as a mobile_element in the
611 submitted chromosome and or plasmid(s) of each strain.
612
613 ***Insertion sequences***
614 Insertion sequences in the plasmids as well as surrounding the CS2 loci located on the
615 chromosome of E1649 were annotated using both Galileo AMR software [91] and the
616 ISFinder database [92]. Complete and partial IS elements were annotated (>95%
617 identity with hits in ISFinder) along with the present genes encoding transposases.
618 Three new insertion sequences were detected in this analysis and were submitted to
619 ISFinder as TnEc2, TnEc3 and TnEc4. Transposons and other mobile elements

620 (integrons and group II introns) were also identified using Galileo AMR and blastn
621 against public databases.
622
623
624 **Supplementary information**
625 **Additional file 1:** Supplemental Figures and Tables
626 **Additional file 2:** Detailed description of ETEC reference plasmids
627 **Additional file 3:** Excel file with plasmid classification – Inc groups
628 **Additional file 4:** Excel file with metadata of ETEC and *E. coli* genomes included in
629 the phylogenetic tree
630 **Additional file 5:** Excel file with a compiled list of all accession numbers for
631 chromosomes and plasmids
632
633 **Acknowledgements**
634 No acknowledgements to mention.
635
636 **Authors' contributions**
637 AvM conceived and designed the experiments, performed the experiments, analysed
638 the data, contributed reagents/materials/analysis tools, prepared figures and/or tables,
639 authored the paper and approved the final draft.
640 GB and AvM annotated all IS elements, transposons as well as other mobile elements,
641 contributed to the paper and approved the final draft.
642 DP and AvM annotated the identified prophages, contributed to the paper and approved
643 the final draft.

644 CB performed the *in silico* analysis of the genomic antibiotic resistance profiling,
645 contributed to the paper and approved the final draft.
646 EJ performed the antibiotic resistance profiling, contributed to the paper and approved
647 the final draft.
648 AJP assembled the genomes, contributed to the paper and approved the final draft.
649 AMS conceived and designed the experiments contributed to the paper and approved
650 the final draft.
651 GD conceived and designed the experiments and approved the final draft.
652 ÅS conceived and designed the experiments, analysed data, authored the paper and
653 approved the final draft.
654

655 **Funding**
656 AvM, AMS and ÅS were supported by the Swedish Foundation for Strategic Research
657 (grant nr. SB12-0072). AvM was also supported by The Swedish Research Council
658 (grant nr. 2018-06828) and the Swedish Society for Medical Research (P18-0140). AJP
659 was supported by the Biotechnology and Biological Sciences Research Council
660 (BBSRC); this research was funded by the BBSRC Institute Strategic Programme
661 Microbes in the Food Chain BB/R012504/1. GD was supported by the Wellcome Trust
662 (grant WT 098051).
663

664 **Availability of data**
665 The datasets supporting the conclusions of this article are included within the articles
666 and its additional files. The sequencing data generated in this study has been
667 submitted to EMBL (Additional file 4 and 5). The physical ETEC reference strains
668 can be requested by contacting the corresponding author Astrid von Mentzer

669 (avm@sanger.ac.uk or mentzerv@chalmers.se). The database used for annotating
670 ETEC virulence factors, ETEC virulence database, including the LT and ST alleles
671 can be found in the github repositories: https://github.com/avonm/ETEC_vir_db and
672 https://github.com/avonm/ETEC_toxin_variants_db.
673 An interactive version of the core genome phylogeny of the 1,065 *E. coli* and ETEC
674 isolates along with the ETEC reference strains (Figure S10) reported here is
675 accessible at https://microreact.org/project/2ZZzaHzeXbMEw9U2MAk7pK?tt=cr
676

677 **Availability of material**

678 Requests for obtaining clinical isolates collected as part of this study should be
679 addressed to the corresponding author. Exchange of clinical isolates should always be
680 in agreement with the University of Gothenburg.
681

682 **Ethics approval and consent to participate**

683 Not applicable.
684

685 **Consent for publication**

686 Not applicable.
687

688 **Competing interests**

689 The authors declare that there are no conflicts of interest.
690

691 **Author details**

692 [1]Department of Microbiology and Immunology, Sahlgrenska Academy, University of
693 Gothenburg, Medicinaregatan 7, 413 90 Gothenburg, Sweden

694 [2]Wellcome Sanger Institute, Hinxton, CB10 1SA, Cambridge, UK

695 [3]EMBL-EBI, Hinxton, CB10 1SA, Cambridge, UK

696 [4] University of Cambridge Department of Medicine, Addenbrookes Hospital

697 Box 157, Hills Rd, CB2 0QQ, Cambridge, UK

698 [5] Department of Microbiology, Tumor and Cell Biology, Biomedicum, Karolinska

699 Institutet, Solnavägen 9, 171 65 Solna, Stockholm, Sweden

700 [6]Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ Norwich, UK

701 [#]Currently affiliated with Chalmers University of Technology, Chalmers Tvärgata 3,

702 413 58 Gothenburg, Sweden

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719 **References**

720 1. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity
721 and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the
722 Global Burden of Disease Study 1990-2016. The Lancet Infectious diseases.
723 2018;18:1229–40.

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

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

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

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

734 6. Harro C, Bourgeois AL, Sack D, Walker R, DeNearing B, Brubaker J, et al. Live
735 attenuated enterotoxigenic Escherichia coli (ETEC) vaccine with dmLT adjuvant
736 protects human volunteers against virulent experimental ETEC challenge. Vaccine.
737 2019;37:1978–86.

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

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

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

747 10. Nada RA, Shaheen HI, Khalil SB, Mansour A, El-Sayed N, Touni I, et al.
748 Discovery and phylogenetic analysis of novel members of class b enterotoxigenic
749 Escherichia coli adhesive fimbriae. Journal of clinical microbiology. 2011;49:1403–
750 10.

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

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

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

761 14. Pichel M, Binsztein N, Viboud G. CS22, a novel human enterotoxigenic
762 Escherichia coli adhesin, is related to CS15. Infection and immunity. 2000;68:3280–
763 5.

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

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

772 17. Joffre E, von Mentzer A, Svennerholm A-M, Sjöling A. Identification of new
773 heat-stable (STa) enterotoxin allele variants produced by human enterotoxigenic
774 Escherichia coli (ETEC). International journal of medical microbiology : IJMM.
775 2016;306:586–94.

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

779 19. Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M, Cunningham
780 AF, et al. A commensal gone bad: complete genome sequence of the prototypical
781 enterotoxigenic Escherichia coli strain H10407. Journal of bacteriology.
782 2010;192:5822–31.

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

786 21. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, et al. The
787 pangenome structure of Escherichia coli: comparative genomic analysis of E. coli
788 commensal and pathogenic isolates. Journal of bacteriology. 2008;190:6881–93.

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

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

796 24. Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, et al.
797 A comparative genomic analysis of diverse clonal types of enterotoxigenic
798 Escherichia coli reveals pathovar-specific conservation. Payne SM, editor. Infection
799 and immunity. 2011;79:950–60.

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

802 26. Sahl JW, Sistrunk JR, Fraser CM, Hine E, Baby N, Begum Y, et al. Examination
803 of the Enterotoxigenic Escherichia coli Population Structure during Human Infection.
804 mBio. 2015;6:e00501.

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

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

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

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

817 31. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the
818 Escherichia coli phylogenetic group. Applied and environmental microbiology.
819 2000;66:4555–8.

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

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

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

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

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

835 37. Fleckenstein JM, Roy K, Fischer JF, Burkitt M. Identification of a two-partner
836 secretion locus of enterotoxigenic Escherichia coli. Infection and immunity.
837 2006;74:2245–58.

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

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

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

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

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

855 43. Sjöling Å, Wiklund G, Savarino SJ, Cohen DI, Svennerholm A-M. Comparative
856 analyses of phenotypic and genotypic methods for detection of enterotoxigenic
857 Escherichia coli toxins and colonization factors. Journal of clinical microbiology.
858 2007;45:3295–301.

859 44. Wachsmuth K, Wells J, Shipley P, Ryder R. Heat-labile enterotoxin production in
860 isolates from a shipboard outbreak of human diarrheal illness. Infect Immun.
861 1979;24:793–7.

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

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

869 47. EUCAST EC for AST. Determination of minimum inhibitory concentrations
870 (MICs) of antibacterial agents by broth dilution. John Wiley & Sons, Ltd (10.1111);
871 2003 Aug p. ix–xv.

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

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

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

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

883 52. Colavecchio A, Jeukens J, Freschi L, Rheault J-GE, Kukavica-Ibrulj I, Levesque
884 RC, et al. Complete Genome Sequences of Two Phage-Like Plasmids Carrying the
885 CTX-M-15 Extended-Spectrum β-Lactamase Gene. Genome announcements.
886 2017;5:90.

887 53. Jacoby GA. Mechanisms of Resistance to Quinolones. Clin Infect Dis.
888 2005;41:S120–6.

889 54. Gilcrease EB, Casjens SR. The genome sequence of Escherichia coli tailed phage
890 D6 and the diversity of Enterobacteriales circular plasmid prophages. Virology.
891 2018;515:203–14.

892 55. Beatty ME, Bopp CA, Wells JG, Greene KD, Puhr ND, Mintz ED. Enterotoxin-
893 producing Escherichia coliO169:H41, United States. Emerging infectious diseases.
894 2004;10:518–21.

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

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

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

904 59. Gómez-Duarte OG, Kaper JB. A plasmid-encoded regulatory region activates
905 chromosomal eaeA expression in enteropathogenic Escherichia coli. Infect Immun.
906 1995;63:1767–76.

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

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

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

915 63. Nishikawa Y, Helander A, OGASAWARA J, MOYER NP, HANAOKA M,
916 HASE A, et al. Epidemiology and properties of heat-stable enterotoxin-producing
917 Escherichia coliserotype O169[ratio]H41. Epidemiology and infection. 1998;121:31–
918 42.

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

923 65. Sack DA, Shimko J, Torres O, Bourgeois AL, Francia DS, Gustafsson B, et al.
924 Randomised, double-blind, safety and efficacy of a killed oral vaccine for
925 enterotoxigenic E. Coli diarrhoea of travellers to Guatemala and Mexico. Vaccine.
926 2007;25:4392–400.

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

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

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

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

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

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

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

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

950 74. Kuhlmann FM, Martin J, Hazen TH, Vickers TJ, Pashos M, Okhuysen PC, et al.
951 Conservation and global distribution of non-canonical antigens in Enterotoxigenic
952 Escherichia coli. Torres AG, editor. PLoS neglected tropical diseases.
953 2019;13:e0007825.

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

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

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

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

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

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

970 81. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.
971 2014;30:2068–9.

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

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

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

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

982 86. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD
983 2017: expansion and model-centric curation of the comprehensive antibiotic
984 resistance database. Nucleic acids research. 2017;45:D566–73.

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

988 88. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al.
989 Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother.
990 2012;67:2640–4.

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

993 90. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better,
994 faster version of the PHAST phage search tool. Nucleic acids research. 2016;44:W16-
995 21.

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

999 92. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference
1000 centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34:D32–6.

1001

1002

1003

1004

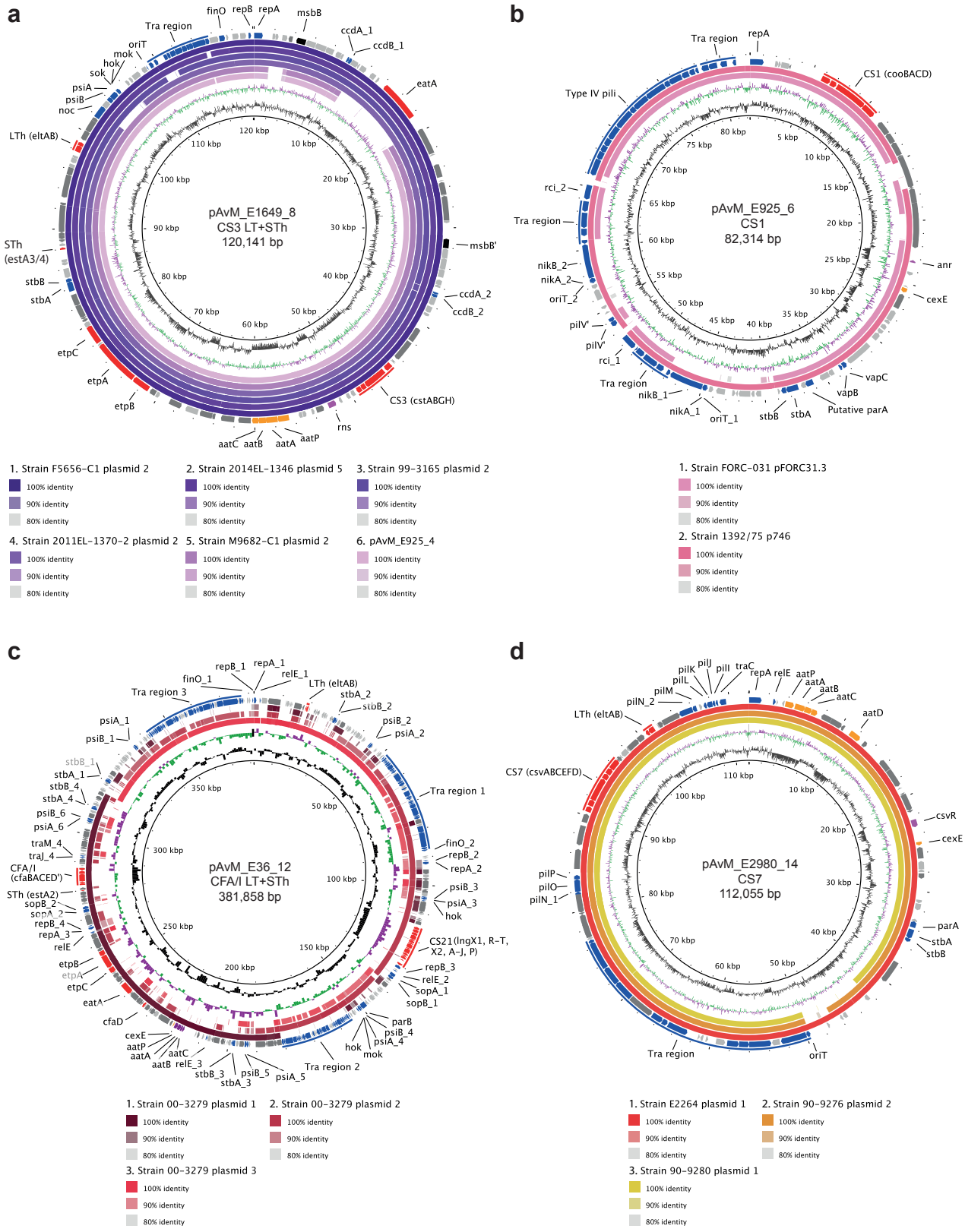
1005

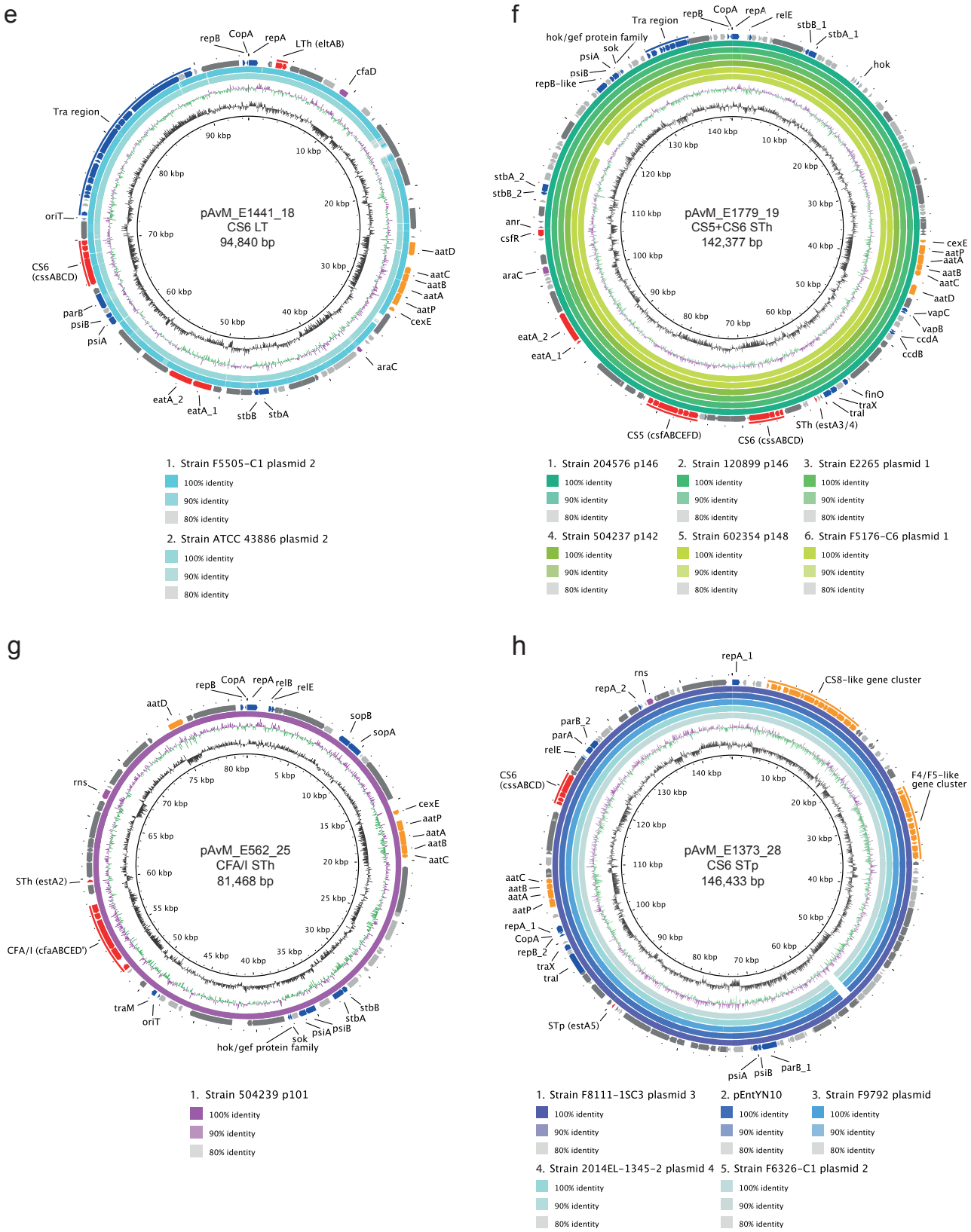
1006

1007

1008

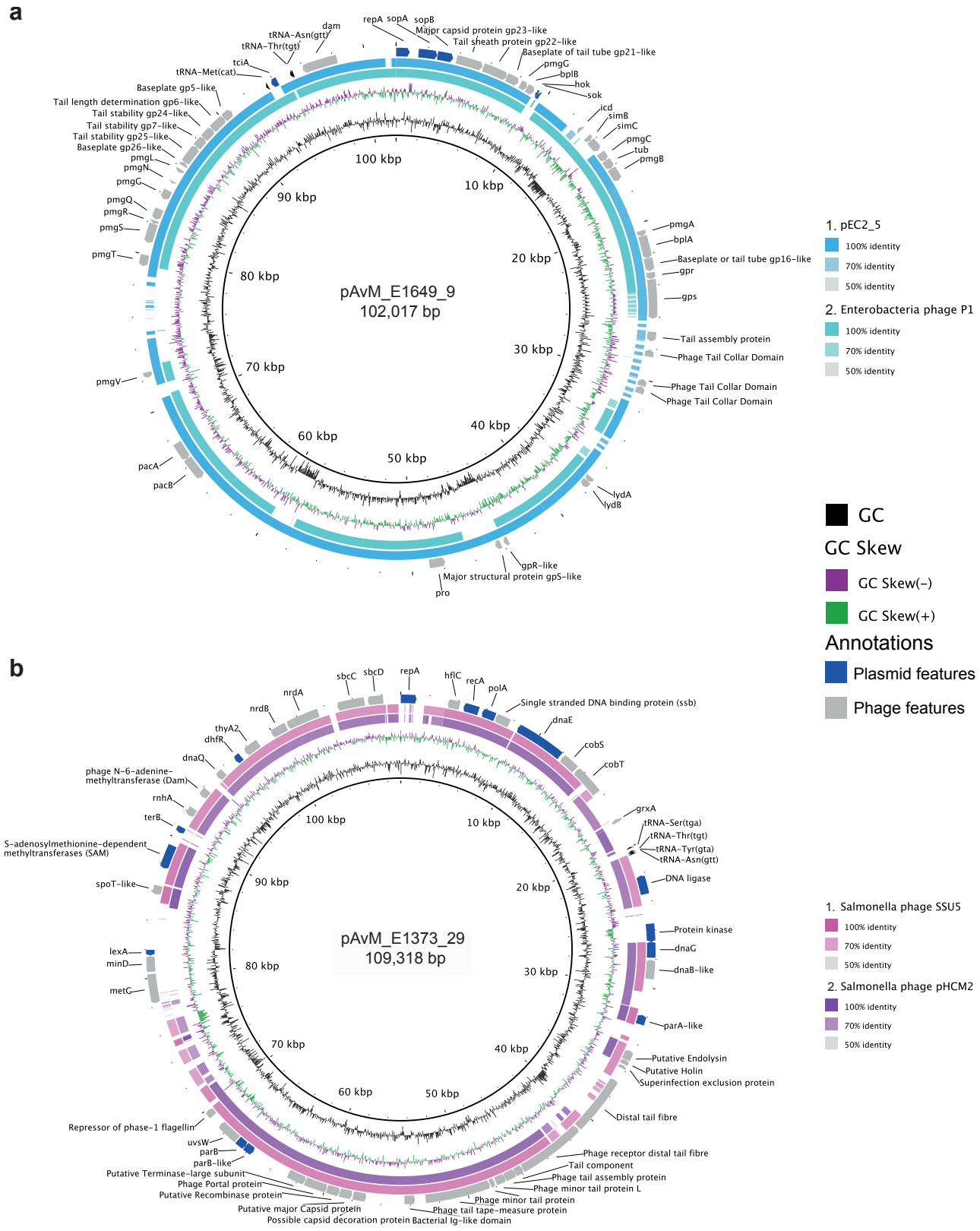
1009


a
pAvM_E1649_8
CS3 LT+STh
120,141 bp
finO
repB
repA
msbB
Tra region
ccdA_1
ccdB_1
oriT
mok
hok
sok
psiA
psiB
noc
eatA
LTh (eltAB)
msbB'
STh (estA3/4)
ccdA_2
ccdB_2
stbB
stbA
etpC
etpA
etpB
CS3 (cstABGH)
rns
aatC
aatB
aatA
aatP
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
Tra region
nikB_2
nikA_2
oriT_2
pilV'
pilV'
rci_1
Tra region
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
100% identity
90% identity
80% identity
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
psiB_5
stbA_3
psiA_5
Tra region 2
hok
mok
psiA_4
psiB_4
parB
sopB_1
sopA_1
relE_2
repB_3
CS21 (lngX1, R-T, X2, A-J, P)
psiA_3
psiB_3
hok
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
pilL
pilM
pilN_2
LTh (eltAB)
aatD
CS7 (csvABCEFD)
csvR
cexE
pilP
pilO
pilN_1
parA
stbA
stbB
Tra region
oriT
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
pAvM_E1441_18
CS6 LT
94,840 bp
repB
CopA
repA
LTh (eltAB)
cfaD
Tra region
oriT
CS6 (cssABCD)
parB
psiB
psiA
eatA_2
eatA_1
stbB
stbA
araC
cexE
aatP
aatA
aatB
aatC
aatD
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
pAvM_E1779_19
CS5+CS6 STh
142,377 bp
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
CS5 (csfABCEFD)
CS6 (cssABCD)
STh (estA3/4)
finO
traX
traI
ccdB
ccdA
vapB
vapC
aatD
aatC
aatB
aatA
aatP
cexE
1. Strain 204576 p146
100% identity
90% identity
80% identity
2. Strain 120899 p146
100% identity
90% identity
80% identity
3. Strain E2265 plasmid 1
100% identity
90% identity
80% identity
4. Strain 504237 p142
100% identity
90% identity
80% identity
5. Strain 602354 p148
100% identity
90% identity
80% identity
6. Strain F5176-C6 plasmid 1
100% identity
90% identity
80% identity
g
pAvM_E562_25
CFA/I STh
81,468 bp
repB
CopA
repA
relB
relE
aatD
sopB
sopA
rns
STh (estA2)
CFA/I (cfaABCED')
traM
oriT
hok/gef protein family
sok
psiA
psiB
stbA
stbB
cexE
aatP
aatA
aatB
aatC
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
1. Strain F8111-1SC3 plasmid 3
100% identity
90% identity
80% identity
2. pEntYN10
100% identity
90% identity
80% identity
3. Strain F9792 plasmid
100% identity
90% identity
80% identity
4. Strain 2014EL-1345-2 plasmid 4
100% identity
90% identity
80% identity
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

1010 **Figure 1:** Comparison between the ETEC reference plasmids harbouring colonisation
1011 factors and other PacBio-sequenced ETEC plasmids using blastn.
1012 **a)** pAvM_E1649_8 (CS3) as reference and pAvM_925_4 (CS3) compared to the
1013 following ETEC plasmids: F5656-C1 plasmid 2 (USA, CP024262.1), 2014-EL-1346-
1014 6 plasmid 5 (2014, USA; CP024237.1), 99-3165 plasmid 2 (USA; CP029980.1),
1015 2011EL-1370-2 plasmid 2 (2011, USA; CP022914.1) and M9682-C1 plasmid 2 (1975,
1016 USA; CP024277.1). **b)** pAvM_E925_6 (CS1) compared to ETEC plasmids
1017 pFORC31.3 (2004, Korea; CP013193.1) and 1392/75 p746 (1973; FN822748.1). **c)**
1018 pAvM_E36_12 (CFA/I) compared to plasmids 1-3 (p1: CP024294.1; p2 CP024295.1;
1019 p3: CP024296.1) from ETEC strain 00-3279 (USA). **d)** pAvM_E2980_14 (CS7)
1020 compared to E2264 plasmid 1 (2006, Bangladesh; CP023350.1), 90-9276 plasmid 2
1021 (1988, Bangladesh; CP024298.1) and 90-9280 plasmid 1 (1988, Bangladesh;
1022 CP024241.1). **e)** pAvM_E1441_18 (CS6) compared to F5505-C1 plasmid 2 (2013,
1023 Sweden; CP023259.1) and ATCC 43886 plasmid 2 (CP024255.1). **f)**
1024 pAvM_E1779_19 (CS5+CS6) compared to 204576 p146 (2010, Mali; CP025908.1),
1025 120899 p146 (2012, Gambia; CP025917.1), E2265 plasmid 1 (2006, Bangladesh;
1026 CP023347.1), 504237 p142 (2010, India; CP025863.1), 602354 p148 (2009,
1027 Bangladesh; CP025848.1) and F5176-C6 plasmid 1 (1997; CP024668.1). **g)**
1028 pAvM_E562_25 (CFA/I) compared to p504239_101 (2010, India; CP025860.1). **h)**
1029 pAvM_E1373_28 (CS6) compared to F8111-1SC3 plasmid 3 (USA; CP024272.1),
1030 pEntYN10 (1991, Japan; AP014654.2), F9792 plasmid (USA; CP023274.1), 2014EL-
1031 1345-2 plasmid 4 (2014, USA; CP024227.1) and F6326-C1 plasmid 2 (1998, USA;
1032 CP024265.1). The thresholds chosen for the blastn are shown in the key below each
1033 plasmid comparison. The colour code for the annotations are listed at the bottom of the

1034 figure. The two most inner rings depict GC content in black and GC Skew- in purple
1035 and GC Skew+ in green. The figures were generated using BRIG [84] v0.95.
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057


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
pacA
pacB
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)
terB
rnhA
phage N-6-adenine-methyltransferase (Dam)
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

1058 **Figure 2:** Comparisons between the identified ETEC phage-plasmids and other similar
1059 phage-plasmids using blastn. **a)** pAvM_E1649_9 is a P1-like phage-plasmid here
1060 compared to Enterobacteria phage P1 (Escherichia virus P1; NC_005856.1) and
1061 pEC2_5 (*E. coli* strain EC2_5; CP041960.1). **b)** pAvM_E2980_29, a phage-plasmid
1062 similar to the pHCM2 (Salmonella typhi strain CT18; AL513384.1) and SSU5
1063 (Salmonella phage; JQ965645.1) in Salmonella typhi. Blastn comparisons were made
1064 using BRIG [84] (v0.95) with the thresholds indicated to the right of each plasmid
1065 comparison. Selected phage and plasmid annotations are shown in the outer ring.
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076

Table 1: Characteristics of the reference ETEC strains.

Strain	Lineage	Phylo-group	MLST	O antigen	CF	Toxin profile	YoI[a]	Location	Subject	Age of subject	D/AS[b]
E925	L1	A	2353	O6	CS1+CS3+CS21	LT+STh	2003	Guatemala	Indigenous	Child <5 yrs	D
E1649	L2	A	4	O6	CS2+CS3+CS21	LT+STh	1997	Indonesia	Traveller	Adult	D
E36	L3	B1	173	O78	CFA/I+CS21	LT+STh	1980	Bangladesh	Indigenous	Adult/child	D
E2980	L3	B1	5305	O114	CS7	LT	2010	Bangladesh	Indigenous	Child <5 yrs	D
E1441	L4	A	1312	O25	CS6+CS21	LT	1997	Kenya	Traveller	Adult	D
E1779	L5	B1	443	O115	CS5+CS6	LT+STh	2005	Bangladesh	Indigenous	Adult	D
E562	L6	A	2332	ON3	CFA/I+CS21	STh	2000	Mexico	Traveller	Adult	D/AS
E1373	L7	E	182	O169	CS6	STp	1996	Indonesia	Traveller	Adult	D

[a] YoI: Year of isolation
[b] D/AS: Diarrhoea or Asymptomatic

Table 2: Description of the plasmids present in the 8 ETEC reference strains.

	Plasmid	**Length (bp)**	**GC (%)**	**Inc[a]**	**Plasmid features (no of copies)**	**Virulence genes**	**Putative virulence genes**	**Antibiotic resistance profile (genomic)**	**Acc. no**
L1 **E925** **CS1+CS3+CS21** **LTh+STh**	pAvM_E925_4	116 803	48.4	FII	*ccdAB, hok-sok* (antisense RNA-regulated system), *psiAB, repAB, stbAB, tra* genes	*cstA-G* (CS3), *eltAB1* (LTh), *estA3/4* (STh)	*etpBAC*	-	LR883051
	pAvM_E925_5	82 909	48.6	FII + FIB	*psiAB, repA* (2), *repB, repE, parB, sopAB, tra* genes	*lngX1, R, S, T, X2, A-J, P* (CS21)	-	-	LR883052
	pAvM_E925_6	82 314	47.8	I1	*iib* (colicin 1b), *repA, stbAB, tra* genes, *pil* locus, *vapBC*	*cooB, A, C, D* (CS1)	*cexE*	-	LR883053
	pAvM_E925_7	51 418	45.4	FII	*ccdAB, psiA, repAB, stbAB, tra* genes	-	*eatA_1-5* (two disrupted *eatA* copies)	-	LR883054
L2 **E1649** **CS2*+CS3+CS21** **LTh+STh**	pAvM_E1649_8	120 141	47.2	FII	*ccdAB* (duplicated), *hok-sok* (antisense RNA-regulated system*), psiAB, repAB, stbAB, tra* genes	*cstA-G* (CS3), *eltAB1* (LTh), *estA3/4* (STh)	*eatA, etpBAC*	-	LR882976
	pAvM_E1649_9*	102 017	47.6	Y	P1 addiction system (phage related), *repA, sopAB*		-	-	LR882977
	pAvM_E1649_10	86 517	45.0	FII+FIB	*hok-sok* (antisense RNA-regulated system), *psiAB, repA*(2)*, repB, sopAB, tra* genes	*lngX1, R, S, T, X2, A-J, P* (CS21)	-	-	LR882974
	pAvM_E1649_11 *	8 834	42.9	No hits	ND	-	-	-	LR882975
L3 **E36** **CFA/I+CS21** **LTh+STh**	pAvM_E36_12*	381 858	49.7	FII + FIB	*stbAB* (4), *psiAB* (6), *tra* genes, r*epA* (3), *repB* (3), *sopAB* (2), *hok-sok* (antisense RNA-regulated system), *relE*	*cfaA, B, C, D* (CFA/I), *eltAB15* (LTh), *lngX1, R, S, T, X2, A-J, P* (CS21), *estA2* (STh)	*eatA, etpBAC*	-	LR882998
	pAvM_E36_13	99 448	51.6	B/O/K/Z	*stbAB, relE, repA, pil* genes, *psiAB*	-	-	*tetA, B, C, R mdf(A)-like*	LR882999
L3 **E2980** **CS7** **LTh**	pAvM_E2980_14	112 056	48.1	I1	*parA, pil* genes, *relE, repA, stbAB*	*csvA, B, C, D* (CS7), *eltAB* (LTh)	*cexE*	-	LR882979
	pAvM_E2980_15	72 255	52.4	FII	*psiAB, relE, repAB, stbAB, tra* genes	-	-	*strA, strB, sul2, blaTEM-1B*	LR882980
	pAvM_E2980_16	48 305	50.3	I1-like	*stbAB, repA, vapBC* (TA-system)	-	*eatA, etpBAC*	-	LR882981

L4 **E1441** **CS6+CS21** **LTh (LT17)**	pAvM_E1441_17	130 302	51.3	FII + FIB	*pemI/K* (TA-system), *psiAB, repA* (2), *repB, sopAB, srnAC* (antisense RNA-regulated system), *stbAB, tra* genes	*lngX1, R, S, T, X2, A-J, P* (CS21)	-	*aadA1, tetR, tetA, sul1, dfrh1*	LR883013
	pAvM_E1441_18	94 840	47.1	FII	*parB, psiAB, repAB, stbAB*	*cssA, B, C, D* (CS6), *eltAB17* (LTh)	*etaA_1, eatA_2, cexE*	-	LR883014
L5 **E1779** **CS5+CS6** **LTh+STh**	pAvM_E1779_19	142 377	47.6	FII	*ccdAB* (TA-system), *cea/cia* (Colicin E), *psiAB, repAB, stbAB* (2 copies), *tra* genes, *vapBC* (TA-system)	*csfA, B, C, E, F, D* (CS5), *cssA, B, C, D* (CS6), *estA3/4* (STh)	*eatA_1, eatA_2*	-	LR883008
	pAvM_E1779_20	88 759	51.8	FII	*hok-sok* (antisense RNA-regulated system), *psiAB, tra* genes, *repAB, stbAB*	*eltAB15* (LTh)	-	-	LR883009
	pAvM_E1779_21	82 464	51.0	FIIY	*repA* (2), *repB, tra* genes, *psiAB, parB, sopAB*	-	-	-	LR883010
	pAvM_E1779_22	61 528	50.5	FII	*repAB, tra* genes	-	-		LR883011
L6 **E562** **CFA/I+CS21** **STh**	pAvM_E562_23	109 853	50.5	I1 (+ FII)	*parB, psiAB, stbAB, pil* genes, *tra* genes	-	*eatA, etpBAC*	-	LR883001
	pAvM_E562_24	86 655	48.6	FII + FIB	*psiAB, repA* (2), *repB, stbAB, tra* genes	*lngX1, R, S, T, X2, A-J, P* (CS21)	-	-	LR883002
	pAvM_E562_25*	81 468	46.7	FII	*psiAB* (truncated *psiA*) *relE/B* (toxin-antitoxin system), *repAB, stbAB, sopAB*	*cfaA, B, C, D* (CFA/I), *estA2* (STh)	-	-	LR883003
	pAvM_E562_26	88 318	52.9	B/O/K/Z	*pil* genes, *pndAC* (antisense RNA-regulated system) *psiAB, relE, repA, tra* genes	-	-	-	LR883004
	pAvM_E562_27*	83 375	40.0	FII	*hok-sok* (antisense RNA-regulated system), *pemI/pemK* (TA system), *psiAB, parB, repAB, tra* genes	-	-	*blaTEM-1b, tetAR, merRTPCADE* (Tn21)	LR883005
L7 **E1373** **CS6** **STp**	pAvM_E1373_28	146 433	46.1	FII + FIB	*parB, psiAB, relE, repA* (2), *repB*	*cssA, B, C, D* (CS6), *estA5* (STp), CS8-like gene cluster, *fae*-related genes	-	-	LR882991
	pAvM_E1373_29	109 318	46.4	FIB	*parB, parB-like, repA*	-	-	-	LR882992

[a]Incompatibility groups were determined by PlasmidFinder [34] and pMLST [34] for subtyping.
*Plasmids not circularised