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 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 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 world-wide.

Methods: Here we have used PacBio long read sequencing in combination with Illumina sequencing to create high quality complete reference genomes for each of these 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.

Results: The ETEC reference strains harbour between two and five plasmids, including virulence, antibiotic resistance and phage-plasmids. The virulence plasmids carrying the colonization 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 superiority as pathogens.

Conclusion: We confirm that the major ETEC lineages all harbor conserved plasmids that have been associated to their respective background genomes for decades. The in-depth analysis of gene content and order and correct annotations of plasmids will help to elucidate other plasmids with and without virulence factors in related bacterial species. These reference genomes allow for rapid 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

50 Diarrheal pathogens are a leading cause of morbidity and mortality globally (WHO,
51 2017), with enterotoxigenic *Escherichia coli* (ETEC) accounting for a large proportion
52 of the diarrhoea cases in resource-poor countries [1]. An estimation of 220 million cases
53 each year are estimated to be attributed to ETEC (WHO PPC 2020). The most
54 vulnerable group are children under five years, but ETEC can also cause disease in
55 adults and is the major cause of diarrhoea in travellers. Resource-poor settings, where
56 access to clean water is limited, enables the spread of ETEC which is transmitted via
57 the faecal-oral route through ingestion of contaminated food or water [2]. The disease
58 severity may range from mild to cholera-like symptoms with profuse watery diarrhoea.
59 The infection is usually self-limiting, lasting three to four days and may be treated by
60 water and electrolyte rehydration to balance the loss of fluids and ions. There is strong
61 evidence to support that an ETEC vaccine is of key importance to prevent children and
62 adults from developing ETEC disease [3]. Several efforts are on-going to develop an
63 ETEC vaccine, with the majority focusing on including immunogenic antigens possibly
64 capable of inducing protection against a majority of the circulating ETEC clones [4–6].
65
66 ETEC bacteria adhere to the small intestine through fimbrial, fibrillar or afimbrial outer
67 membrane-structures called colonisation factors (CF). Upon colonisation, the bacteria
68 proliferate and secrete heat-labile toxin (LT) and/or heat-stable toxins, (STh or STp)
69 causing diarrhoea and often vomiting allowing further spread of the bacteria in the
70 environment [7].
71
72 The ability of an ETEC strain to infect correlates with its ability to adhere to cells of a
73 specific host. To this day 27 different CFs with human tropism have been described
74 and individual ETEC strains usually express 1-3 different CFs [8–14]. The

75 enterotoxins, LT and ST, can also be subdivided based on structure and function.
76 Human-associated ETEC strains express one of the 28 different LT-I variants (LTh-1
77 and LTh-2 are the most common variants) [15] alone or together with one of the
78 genetically different types of STa; STh and STp [16,17].
79
80 We have previously shown that ETEC strains causing human disease can be grouped
81 into a set of clonal lineages that encompass strains with specific virulence profiles.
82 Seven of the 21 identified lineages encompass ETEC strains that express the most
83 commonly found CFs and toxin profiles amongst isolated clinical ETEC strains [4,18].
84
85 There is currently one complete ETEC reference genome, H10407 [19], which has been
86 manually annotated. Several additional complete ETEC genomes are available [20,21],
87 some of which are annotated using automated annotation pipelines that often fail at
88 correctly annotating ETEC specific genes such as CFs. The rapid adaptation of next-
89 generation sequencing in public health, specifically within bacterial diseases [22,23]
90 and a number of large-scale sequencing studies [18,24–28] has led to a rapid increase
91 in the number of publicly available ETEC genomes. Most of these data were generated
92 with short-read technologies, such as Illumina. A limitation of short-read sequence data
93 is the inability to unambiguously resolve short repetitive regions of a genome, leading
94 to fragmented de novo assemblies of the underlying genome, missing regions and
95 genes, and disjointed synteny. ETEC is a highly diverse pathogen both in the core
96 genome (SNPs) and the accessory genome (mobile genetic elements). Clinically
97 important mobile genetic elements, such as virulence plasmids, vary within ETEC
98 strains. Hence, it is important to identify lineage-specific reference genomes that are
99 fully annotated and include both chromosome and plasmid(s). A number of fully

100 complete genomes have been generated using long-read sequencing alone [20,28],
101 however circularising some chromosomes and plasmids may be difficult due to repeat
102 regions. This can be resolved using a hybrid assembly approach combining long-read
103 and short-read sequencing data. In this report we describe eight fully complete
104 annotated genomes and plasmids from isolates representing the major ETEC lineages
105 (L1-L7) that cause disease globally. They are sequenced using both short and long-read
106 sequencing technologies to provide the highest accuracy currently available. These
107 reference genomes will form the foundation of ETEC genomics research for years to
108 come.
109

110 **Methods**

111 ***Selection of strains***

112 Initially one to two ETEC strains with in each of the lineage (L1-L7)-specific CF profile
113 were chosen from the University of Gothenburg large collection of ETEC strains [18]
114 for PacBio sequencing. The strains were selected based on the location and year of
115 isolation with the aim to select strains isolated from patients with diarrhoea from
116 diverse geographical locations and at different time-points. After the genomes had been
117 sequenced, assembled, circularised and annotated a second selection was made for
118 manual curation of the genomes. This selection was made based on the quality of the
119 genome assembly and the circularisation. The whole genomes of the ETEC reference
120 strains was compared with one or two other long-read sequenced ETEC strains
121 belonging to the same lineage by progressiveMAUVE and show that the strains are
122 colinear (Figure S1). One representative ETEC genome from each lineage was
123 annotated, with emphasis on the plasmids. The physical ETEC reference strains are
124 available upon request.

125

126 ***Phenotypic toxin and CF analyses***

127 ETEC isolates were identified by culture on MacConkey agar followed by analysis of
128 LT and ST toxin expression using GM1 ELISAs [29]. The expression of the different
129 CFs was confirmed by dot-blot analysis [29]. Isolates had been kept in glycerol stocks
130 at −70 °C, and each strain has been passaged as few times as possible.

131

132 ***Antibiotic susceptibility testing***

133 All ETEC isolates were tested against 14 antimicrobial agents and their minimum
134 inhibitory concentration was determined by broth microdilution using EUCAST
135 methodology [30]. The antimicrobial agents were: ampicillin, amoxicillin-clavulanic,
136 oxacillin, ceftazidime, ceftriaxone, doxycycline, tetracycline, nalidixic acid,
137 norfloxacin, azithromycin, erythromycin, chloramphenicol, nitrofurantoin and
138 sulfamethoxazole-trimethoprim. All antibiotics were purchased from Sigma-Aldrich.
139 The *Escherichia coli* ATCC 25922 was used as quality control. The MIC was recorded
140 visually as the lowest concentration of antibiotic that completely inhibits growth.

141

142 ***DNA extraction and sequencing***

143 Strains from each lineage (L1-L7) were SMRT-sequenced on the PacBio RSII. A
144 hybrid *de novo* assembly was performed combining the reads from both the SMRT-
145 sequenced and Illumina sequenced strains.

146 For Single-Molecule Real Time (SMRT) sequencing (Pacific Bioscience) long intact
147 strands of DNA are required. The genomic DNA extraction was performed as follows.
148 Isolates were cultured in CFA broth overnight at 37°C followed by cell suspension in
149 TE buffer (10 mM Tris and 1 mM EDTA pH 8.0) with 25% sucrose (Sigma) followed

150 by lysis using 10 mg/ml lysozyme (in 0.25 Tris pH 8.0) (Roche). Cell membranes were
151 digested with Proteinase K (Roche) and Sarkosyl NL-30 (Sigma) in the presence of
152 EDTA. RNase A (Roche) was added to remove RNA molecules. A phenol-chloroform
153 extraction was performed using a mixture of Phenol:Chloroform:Isoamyl Alcohol
154 (25:24:1) (Sigma) in phase lock tubes (5prime). To precipitate the DNA 2.5 volumes
155 99% ethanol and 0.1 volume 3 M NaAc pH 5.2 was used fol- lowed by re-hydration in
156 10 mM Tris pH 8.0. DNA concentration was measured using NanoDrop
157 spectrophotometer (NanoDrop). On average 10 μg for PacBio sequencing. Library
158 preparation for SMRT sequencing was prepared according to the manufacturers'
159 (Pacific Biosciences) protocol. The DNA was stored in E buffer and sequenced at the
160 Wellcome Sanger Institute. Isolates were sequenced with a single SMRTcell using the
161 P6-C4 chemistry, to a target coverage of 40–60X using the PacBio RSII sequencer.

162 ***Assembly***

163 The resulting raw sequencing data from SMRT sequencing was manually *de novo*
164 assembled using the PacBio SMRT analysis pipeline
165 (https://github.com/PacificBiosciences/SMRT-Analysis) (version 2.3.0) utilizing the
166 Hierarchical Genome Assembly Process (HGAP) [31]. For all samples, the unfinished
167 assembly produced a single, non-circular, chromosome plus some small contigs, some
168 of which were plasmids or unresolved assembly variants. Using Circlator [32] (version
169 1.1.0), small self-contained contigs in the unfinished assembly were identified and
170 removed, with the remaining contigs circularized. Quiver [31] was then used to correct
171 errors in the circularized region by mapping corrected reads back to the circularized
172 assembly. As the strains had also been short read sequenced, and this data is of higher
173 base quality, the short reads from the Illumina sequencing was used in combination
174 with the long reads using Unicycler [33] to generate high quality assemblies (Table S1).

175

176 Fully circularised chromosomes and plasmids were achieved for the majority of the
177 strains. Cross validation of the assemblies was performed where two or three strains of
178 a lineage were sequenced (Figure S1). A single assembly from each lineage was chosen
179 to act as the representative reference genome, with priority given to assemblies with the
180 most complete and circularised chromosome and plasmids. In total, one chromosome
181 and 5 out of the 29 plasmids could not be circularized (independent on the two strains
182 that were sequenced initially) out of the 8 selected representative strains. These are
183 indicated with an asterix in Table 2 and Table S2. Between two and five plasmids were
184 identified in the eight strains. Shorter contigs that could not be assembled properly
185 contained phage genes and are included in the genomes and annotated as prophages
186 Table S5). Socru was used to validate the assembly of the chromosome, they all have
187 biologically valid orientation and order of rRNA operons with a type GS1.0, which is
188 seen in most *E.coli* in the public domain [34].

189

190 ***Gene prediction, annotation and comparative analysis***

191 The final assembly was annotated using Prokka [35] (version 1.14.6). The annotations
192 of all plasmids generated by Prokka were manually checked using the genome viewer
193 Artemis [36] and Geneious 2018.11.1.5 (http://www.geneious.com) together with
194 blastp. Annotations of known ETEC virulence genes (colonisation factors and toxins)
195 and putative virulence genes were added after blast+ analysis against the ETEC
196 virulence database and their annotations updated accordingly. The LT and ST alleles
197 were determined according to Joffre *et al.,* [15,17]. Where required, PFAM domains
198 where searched using jackhammer to back up any identified using blastp
199 (https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer). Blastn and tblastx was used

200 for plasmid comparison, using both NCBI website or within BLAST Ring Image
201 Generator (BRIG) [37].
202
203 ***Incompatibility groups***
204 Due to the discrepancy in databases two approaches was used to determine the Inc
205 groups of the 25 plasmids. PlasmidFinder was used with a threshold for minimum %
206 identity at 95% and a minimum coverage of 60%. The plasmids were further
207 characterised by pMLST [38], except for IncY which are a group of prophages that
208 replicate in a similar manner as autonomous plasmids. IncB/O/K/Z plasmids were
209 further typed by blastn comparison to the reference B/O (M93062), K (M93063) and Z
210 (M93064) replicons.
211
212 ***oriT prediction***
213 The location of the *oriT* in the plasmids, if present, was predicted using oriTFinder [39]
214 with Blast E-value cut-off set to 0.01.
215
216 ***Genomic antibiotic resistance profiling***
217 The identification of antibiotic resistance genes, located on both the chromosome and
218 plasmid(s) as well as the presence of efflux pumps and porins known to confer
219 resistance to antibiotics. The results were obtained by running ARIBA [40] using the
220 CARD database [41] with the default settings (minimum 90% sequence identity and no
221 length cut-off). ARIBA combines a mapping/alignment and targeted local assembly
222 approach to identify AMR genes and variants efficiently and accurately from paired
223 sequencing reads. The heatmaps were visualised using Phandango [42]. The presence

224 of chromosomal mutations in *gyrA* and *parC* was determined with ResFinder v3.2 from
225 the Center of Genomic Epidemiology [43].
226
227 ***Prophage prediction***
228 The complete FASTA sequence of each ETEC reference genome was searched for
229 phage genes and prophages using PHASTER (phaster.ca) [44]. The identified intact
230 prophages are listed in Table S5. Additional questionable and not intact prophages were
231 identified but have not been included here. The prophages have been given a specific
232 identifier name and is also annotated as a mobile_element in the submitted chromosome
233 and or plasmid(s) of each strain.
234
235 ***Insertion sequences***
236 Insertion sequences in the plasmids as well as surrounding the CS2 loci located on the
237 chromosome of E1649 were annotated using both Galileo AMR software [45] and the
238 ISFinder database [46]. Complete and partial IS elements were annotated (>95%
239 identity with hits in ISFinder) along with the present genes encoding transposases.
240 Three new insertion sequences were detected in this analysis and were submitted to
241 ISFinder as TnEc2, TnEc3 and TnEc4. Transposons and other mobile elements
242 (integrons and group II introns) were also identified using Galileo AMR and blastn
243 against public databases.
244
245 **Results**
246 ***Genome analysis of eight representative ETEC isolates***
247 Eight ETEC strains representing the seven major ETEC lineages (L1-L7) comprising
248 isolates with the most prevalent virulence factor profiles were sequenced, assembled,

249 circularised and manually curated (Table 1). L3 includes two different representative
250 strains, one CS7 and one CFA/I positive strain. All chromosomes except one were
251 circularised (E1779). The average length of the chromosome was 4,927,521 bases
252 (4,721,269-5,151,162) with an average GC content of 50.7% (50.4% to 50.9%) and the
253 number of CDS ranging from 4,409 to 4,924 (Table S2). Each ETEC reference genome
254 contains between two and five plasmids encompassing plasmid-specific features,
255 virulence genes (both known and putative virulence factors) as well as antibiotic
256 resistance genes (Table 2).
257
258 ***Comparative genomics of the chromosome***
259 The chromosomes of the reference strains were aligned and compared using MAUVE
260 and the overall structure was conserved across all eight chromosomes (Figure S2). In
261 total, 8,348 chromosomal genes were identified in the eight ETEC strains with 3,179
262 genes considered part of the core genome shared by all eight reference strains. The
263 majority of human commensal *Escherichia coli* strains belong to subgroup A [47,48].
264 ETEC strains however, fall into multiple phylogenetic groups (A, B1, B2, D and E with
265 the majority found in the phylogenetic groups A and B1 [18]. The phylogenetic group
266 of the eight ETEC reference strains have previously been determined using the triplex-
267 PCR scheme [49]. The ETEC references were re-analysed using ClermonTyping [50]
268 and it was determined that strain E1373 belongs to the phylogenetic group E while the
269 other reference isolates belong to groups A and B1 (Table 1).
270
271 ***Plasmids***
272 The plasmids of each isolate were annotated using Prokka followed by manual curation
273 of the annotations including genes part of the conjugation machinery, and known

274 plasmid stability genes. Virulence factors (including CFs, toxins and putative virulence
275 factors), antibiotic resistance determinants with the Comprehensive Antibiotic
276 Resistance Database (CARD) [51]) as well as complete and partial insertion elements
277 and prophages were manually annotated. The plasmids were designated
278 pAvM_strainID_integer, e.g. pAvM_E925_4 (Additional file 1). The first plasmid
279 reported in this study starts at 4 as 3 previous plasmids E873p1-3 already have been
280 deposited to GenBank related to a different project [8].
281
282 Plasmids were typed by analysing the presence and variation of specific replication
283 genes to assign the plasmids to incompatibility (Inc) groups. The Inc groups of the
284 ETEC reference plasmids were first determined using PlasmidFinder and further
285 classified into subtypes using pMLST [38]. The replicons identified are IncFII,
286 IncFIIA, IncFIIS, IncFIB, IncFIC, IncI1 and IncY. Plasmids with replicon IncY,
287 IncFIIY or IncB/O/K/Z mainly harboured plasmid associated genes, such as stability
288 and transfer genes. Importantly, replicons FII, FIB and I1 were found to be strongly
289 associated with virulence genes as all CF locis, toxins as well as putative virulence
290 factors (e.g. *eatA* and *etpBAC*) were present on these plasmids. The majority of all
291 ETEC plasmids analysed here (17/29) belong to IncFII, whereof six of the IncFII
292 plasmids have an additional IncFIB replicon. In 6 of the ETEC reference strains two or
293 three IncFII replicons are present, for example, in strain E925, the plasmids
294 pAvM_E925_4 and 7 both belong to IncFII. However, the plasmids were further
295 subtyped to FII-111 and FII-15, respectively (Table 2) and is likely that their replicons
296 are different enough that they are compatible.
297
298 ***Virulence factors***

299 The CFs expressed by the selected reference strains are CFA/I, CS1-CS3, CS5-CS7 and
300 CS21. Three of the strains (E925, E1649 and E1779) express both LT and ST, two
301 strains (E2980 and E1441) express LT and the strains E36 and E562 express STh, while
302 E1373 express STp (Table 1). Genes encoding virulence genes can be found on the
303 same plasmid, usually a CF locus and genes encoding one or two toxins. Interestingly,
304 genes encoding multiple CFs in the same strain are not commonly located on the same
305 plasmid, but instead found on individual plasmids (in the ETEC reference strains
306 described here). Exceptions for this is strain E1779 in which CS5 and CS6 loci are
307 located on the same plasmid (pAvM_E1779_19). In both E925 (L1) and E1649 (L2)
308 the genes encoding CS3 (*cst*), ST (*estA*) and LT (*eltAB)* are located on the same
309 plasmid, both with the FII replicon and of roughly the same size (Table 2). Blastn
310 comparison between the plasmids as well with additional plasmids, which harbour the
311 same virulence genes, shows that they are highly conserved (Figure 1a). The results
312 correspond with the close genetic relationship and common ancestry of lineage 1 (L1)
313 and lineage 2 (L2) [18].
314 Additional putative virulence factors, besides CFs and toxins, were identified in the
315 majority of the strains (Table 2), with *eatA* and *etpBAC* being the most commonly
316 found. EatA is an immunogenic mucinase which contributes to virulence by degrading
317 MUC2 which is the major protein in mucus of the small intestine [52,53]. The *etpABC*
318 genes encode an adhesin located on the tip of the flagella and mediates adherence to
319 host cells [54,55]. Four of the reference strains (E925, E1649, E36 and E562) harbour
320 both *eatA* and *etpBAC.* In three strains the *eatA* and/or *etpBAC* are located on the same
321 plasmid with an FII or FII+FIB replicon along with additional ETEC virulence genes,
322 except in E562 and E1373, where *eatA* and *etpBA*C are located on an I1+FII
323 (pAvM_E562_23) and I1 (pAvM_E1373_16) plasmid, respectively, which mainly

324 contains plasmid associated genes as well as genes encoding the *pil* operon and *tra-*
325 operon (pAvM_E562_23). Furthermore, a less explored putative virulence factor is
326 CexE, which is an extracytoplasmic protein dependent on the expression of *cfaD,* the
327 CFA/I regulator [56], and was first identified in H10407 [57]. Corroborating earlier
328 findings, the CFA/I positive E36 (L3) and E562 (L6) isolates harbour *cexE*
329 (pAvM_E36_12 and pAvM_E562_25). In addition, *cexE* is present in pAvM_E925_6,
330 pAvM_E1779_19 and pAVM_E2980_14, pAvM_E1441_18 and pAvM_E1373_28.
331 *CexE* has previously also been identified in several CS5+CS6 positive ETEC and
332 shown to be upregulated in the presence of bile and sodium glycocholate-hydrate [58],
333 bile is known to be involved in the regulation of several ETEC CFs [29,59]. The
334 location of *cexE* seems to be conserved across specific strains. In pAvM_E36_12,
335 pAvM_E1441_18, pAvM_E1779_19 and pAvM_E562_25 *cexE* is located upstream of
336 the *aatPABC* locus, whereas in pAvM_E925_6 and pAvM_E2980_14 *cexE* is located
337 downstream of *rob* (an AraC family transcriptional regulator) in the opposite direction.
338 pAvM_E925_4 harbours the aatPABC locus and hence the *cexE* is located on a
339 different plasmid in this strain.
340
341 ***Comparison of plasmids with the same virulence profile***
342 ETEC isolates within a lineage share the same virulence profile, more specifically the
343 same CF profile. Blastn of each of the CF positive plasmids were performed and the
344 best hit(s) were used for subsequent analysis (Figure 1). Most of the plasmids identified
345 as related to the ETEC reference plasmids were not annotated, hence, when needed
346 these were annotated using the corresponding ETEC reference plasmids annotation as
347 a high priority when running Prokka. We show that plasmids with the same virulence
348 profile are rather conserved (Figure 1). For example, the two plasmids encoding CS3

349 (pAvM_E925_4 and pAvM_E1649_8) are highly similar to several CS3 harbouring
350 plasmids from O6:H16 strains collected from various geographical locations between
351 1975 and 2014, including *Escherichia coli* O6:H16 strain M9682-C1 plasmid
352 unnamed2 (CP024277.1) and *Escherichia coli* strain O6:H16 F5656C1 plasmid
353 unnamed2 (CP024262.1) PacBio sequenced by Smith et al. [20] (Figure 1a).
354 Furthermore, high coverage and similarity were found between the plasmids of isolates
355 E1441 (L4) and PacBio sequenced plasmids of ETEC isolates ATCC 43886/E2539C1
356 and 2014EL-1346-6 [20]. These isolates were collected in the 70-ties [60] and 2014
357 (from a CDC collection), respectively, and assigned as O25: H16 which is the O group
358 determined for E1441 *in silico* (Figure 1e). Plasmids of E2980 (LT+ CS7, L3) were
359 validated by PacBio sequenced plasmids of the ETEC isolate E2264, and E1779 (LT,
360 STh+CS5+CS6, L5) were largely corroborated by two plasmids of E2265 (LT,
361 STh+CS5+CS6 [28,58], although four plasmids were found in E1779. Several
362 additional L5 ETEC genomes have been sequenced within the GEMS study [61] and
363 high plasmid similarity and conservation in CS5+CS6 positive L5 isolates was evident
364 (Figure 1).
365
366 Overall the results show that ETEC plasmids are specific to lineages circulating world-
367 wide and conserved over time. Thus, the plasmids of major ETEC lineages must confer
368 evolutionary advantages to their host genomes (Figure 1).
369
370 ***Antibiotic resistance***
371 *Escherichia coli* can become resistant to antibiotics, both via the presence of antibiotic
372 resistance genes as well as the acquisition of adaptive and mutational changes in genes

373 encoding efflux pumps and porins which allows the bacterium to effectively pump out
374 the antibiotic molecules.
375
376 Antibiotic resistance genomic marker(s), both chromosomally located and on plasmids,
377 were identified using the CARD database [51] (Table 2, Figures S4 and Figure S5 and
378 Additional file 1). Similar to other studies IncFII and B/O/K/Z plasmids were found to
379 harbour genes conferring antibiotic resistance [62]. Furthermore, the phenotypic
380 antibiotic resistance profile was determined with clinical MIC breakpoints based on
381 EUCAST (The European Committee on Antimicrobial Susceptibility Testing) [30]
382 (Table S4). Phenotypic antibiotic resistance profiles (Table S4) were largely supported
383 by the findings of antibiotic resistance genes, efflux pumps and porins (Figures S4-S5
384 and Table S3), although some differences were found. All of the ETEC reference strains
385 are phenotypically resistant to at least two antibiotics of the 14 tested (Table S4).
386 Resistance against penicillin's, norfloxacin (Nor) and chloramphenicol (Cm) is most
387 common among these strains. Two of the strains, E1441 and E2980, harbour more than
388 four antibiotic resistance genes as well as multiple efflux systems and porins (Figure
389 S5 and Table S3). The plasmid pAvM_E1441_17 carries *aadA1-like*, *dfrA15*, *sul1* and
390 *tetA*(A) resistance genes (Table 2), where the first three genes are in a Class 1 integron
391 which confers resistance to streptomycin, trimethoprim, and sulphonamide
392 (sulphamethoxazole). The gene *tetA*(A) is part of the truncated Tn*1721* transposon [63].
393 The E1441 strain was verified as resistant to tetracycline (Tet) and sulphamethoxazole-
394 trimethoprim (Sxt) while streptomycin was not tested. A *mer* operon derived from Tn*21*
395 is also present the resistance region of pAvM_E1441_17, indicating that the plasmid
396 would also likely confer tolerance to mercury, although, this was not confirmed (Table
397 2). Interestingly, this multi-replicon (FII and FIB) plasmid also harbours the *lng* locus

398 encoding CS21, one of the most prevalent ETEC CFs. In isolate E2980 virulence
399 plasmid pAvM_E2980_15 harboured multiple resistance genes in the same region
400 (*bla*$_{\text{TEM-1b}}$, *strA*, *strB* and *sul2)* which confer resistance to ampicillin, streptomycin and
401 sulphonamides. E2980 was found to be resistant to ampicillin (Amp) and oxacillin
402 (Oxa), which can be broken down by the beta-lactamase Bla$_{\text{TEM-1b}}$, (Table 2, Table S3
403 and Table S4). E562 harbours three antibiotic resistance genes, *ampC* located in the
404 chromosome and the *tet(A)* and *bla*$_{\text{TEM-1b}}$ genes on a FII plasmid (pAvM_E562_27).
405 The *mer* operon derived from Tn*21* is also present in the region (Table 2 and Table S3).
406 The phenotypic resistance profile of E562 matches the genomic profile with resistance
407 to tetracycline (Tet), ampicillin (Amp), amoxicillin-clavulanic acid (Amc) and oxacillin
408 (Oxa) (Table S4). The plasmid pAvM_E36_13 contains a complete copy of Tn*10*,
409 which encode for the *tet(B)*, tetracycline resistance module. Although the
410 AvM_E1373_29 phage-like plasmid is cryptic, related plasmids such as the pHMC2-
411 family of phage-like plasmids [64] (described below), can harbour resistance genes
412 such as CTX-M-14 [65] and CTX-M-15 [66,67].
413
414 Phenotypic intermediate resistance to ampicillin was found in E36 and E1779 encoded
415 by chromosomal *ampC*. Higher MIC values against ampicillin are found in E2980 and
416 E562 strains carrying *bla*$_{\text{TEM}}$ genes. Phenotypic resistance to ceftazidime (Caz) and
417 ceftriaxone (Cro) was not found in the isolates which was consistent with the absence
418 of extended spectrum beta lactamase (ESBL) resistance genes in the sequence data.
419
420 Phenotypic resistance to norfloxacin (Nor) and chloramphenicol (Cm) was found in 6
421 and 5 isolates, respectively, but none of the resistant isolates contained known

422 resistance genes suggesting that chromosomal mutations or presence of efflux pumps
423 may account for this reduced susceptibility.
424 The presence of efflux systems and porins that can confer resistance to antibiotics and
425 was hence analysed using whole genome sequencing. The ETEC reference strains
426 contain several efflux systems which could explain why the genotypic and phenotypic
427 antibiotic resistance profile did not match for all antibiotics. All of the isolates harbour
428 multiple efflux pumps located on the chromosome and plasmids (Table S3 and Figure
429 S5) and could explain the reduced susceptibility to fluoroquinolones seen in the
430 majority of the reference strains. In E925, a non-synonymous mutation in *acrF* was
431 identified (G1979A) resulting in a substitution from arginine to glutamine (A360Q).
432 The effect on the expression and/or function of the AcrEF efflux pump was not verified.
433
434 In addition, the isolates were analysed for chromosomal mutations likely to confer
435 quinolone resistance, using ResFinder mutations in *gyrA* was only found in one strain,
436 E2980, at position S83A which may confer resistance to nalidixic acid and
437 ciprofloxacin). However, E2980 was sensitive to Nalidixic acid. Both mutation(s) that
438 alter the target (*gyrA* and *parC*) as well as the presence of efflux pumps can confer
439 resistance to fluoroquinolones. The majority of the isolates are moderately resistant to
440 Norfloxacin (and Nalidixic acid), both quinolones, which is most likely due to the
441 presence of two efflux pumps, AcrAB-R and AcrEF-R, as only one mutation was
442 identified in *gyrA* of isolate E2980 where usually at least two or more mutations are
443 needed to confer augmented resistance [68].
444
445 ***Identification of phage-like plasmids in ETEC***

446 Two of the ETEC reference strains (E1649 and E1373) harboured phage-like plasmids
447 (pAvM_E1649_9 and pAvM_E1373_29) which encode for DNA metabolism, DNA
448 biosynthesis as well as structural bacteriophage genes (capsid, tail etc.). Both
449 pAvM_E1649_9 and pAvM_E1373_29 contain genes associated with plasmid
450 replication, division and maintenance (*i.e. repA* and *parAB*). Phage-like plasmids are
451 found in various bacterial species, such as *Escherichia coli*, *Klebsiella pneumoniae*,
452 *Yersinia pestis*, *Salmonella enterica* serovar Typhi, *Salmonella enterica* serovar
453 Typhimurium, *Salmonella enterica* serovar Derby and *Acinetobacter baumanii* [69].
454 The plasmid pAvM_E1649_9 belong to the P1 phage-like plasmid family (Figure 2a
455 and Figure S3a) while pAvM_E1373_29 belongs to the pHCM2-family (Figure 2b and
456 Figure S3b) that can be traced back to a likely phage origin similar to the Salmonella
457 phage, SSU5 [64]. Both phage-plasmids thus contain replication and/or partition genes
458 of plasmid origin and a complete set of genes that are phage related in function and
459 properties (Figure 2 and Figure S3). It is significant that phage-like plasmid
460 pAvM_E1373_29 falls more within the *Escherichia coli* lineage of pHCM2 phage-like
461 plasmid rather than those found in *Salmonella* species. This indicates that phage-like
462 plasmids have diversified within the context of the bacterial species in which they have
463 been isolated.
464
465 Blast searches confirmed high similarity of pAvM_E1373_29 to several phage-like
466 plasmids found in *Escherichia coli* including ETEC O169:H41 isolate F8111-1SC3
467 [20,70], and a number of CTX-M-15 positive phage-like plasmids including pANCO1,
468 pANCO2 [67], PV234a and a plasmid found in *Escherichia coli* ST648 from waste
469 water and ST131 isolate SC367ECC [71]. The P1 phage-like plasmid pAvM_1649_9
470 is most similar to p1107-99K, pEC2_5 isolated from human urine and p2448-3 from an

471 UPEC ST131 isolate isolated from blood. The similarity is most pronounced at the
472 amino acid level. Conservation and synteny is evident when pAvM_1649_9 is
473 compared to P1 phage.
474
475 ***Prophages present and their cargo genes***
476 Prophages may insert into chromosomes and bring along not only the genes required
477 for lysogeny and lytic cycles but also cargo genes that are often picked up when DNA
478 is compacted into the capsid. Cargo genes can have significant benefits to the host
479 bacterium by providing additional elements to defence against phage or immune
480 evasion and finally environmental survival. PHASTER analyses identified prophages
481 in the chromosomes of all ETEC reference isolates and in some of the plasmids (Table
482 S5). Putative tellurite resistance operons in isolates E925, E36, E2980 and E1373 were
483 all located in prohages. In addition, *eatA* (in E925, and E1649) and *estA* (STh) genes
484 (E36) were prophage cargo genes.
485
486 Many prophage cargo genes identified in this study have properties related to inhibition
487 of cell division. Among these are a variety of *kil* genes which can enhance host bacterial
488 survival in the presence of some antibiotics [72]. Some genes that are core entities
489 within many prophages, such as *zapA* (from E1779_Pph_6), *dicB* and *dicC* (found in
490 phage E1779_Pph_7), also have similar effects as they can inhibit cell division in the
491 presence of antibiotics which raise the broader question in terms of how they are
492 beneficial to the host bacterium.
493
494 A further gene of interest is the *yfdR* gene identified in E1779_Pph_7 (gene
495 E1779_04412). This curtails cellular division by inhibiting DNA replication under

496 stress conditions encountered by the bacterial cell. Similarly, the *iraM* gene located in
497 phage E1441_Pph_2 plays a role in RpoS stability.
498
499 OmpX homologs were found in numerous phages in this study. They are trans-
500 membrane located and play a role in virulence as well as antibiotic resistance [73]. PerC
501 is often associated with EPEC plasmids where it seems to have a regulatory role for the
502 attaching and effacing gene, *eaeA* [74]. Its presence as cargo within an ETEC strain
503 phage, E1779_Pph_7, is thus intriguing. Its ability to regulate other virulence genes is
504 yet to be determined. Within the same phage, a *gntR*-like regulatory gene was
505 identified. It play a role in gluconate utilization as well as induction of the genes of the
506 Entner-Doudoroff pathway [75].
507
508 **Discussion**
509 ETEC strains have previously been shown to fall into globally spread genetically
510 conserved lineages which encompass strains with highly specific virulence factor
511 profiles [18]. The currently widely used ETEC reference strains H10407 (CFA/I) and
512 E24377A (CS1+CS3) are highly divergent in relation to strains with the same virulence
513 profile sequenced more recently [18]. This highlights the need of relevant and
514 representative ETEC reference strains and genomes. The long-read sequenced isolates
515 presented here comprise complete reference genomes with separate chromosomal and
516 plasmid sequences that allow more detailed studies of ETEC as well as *Escherichia coli*
517 phylogeny.

518 Phylogenetic analysis has shown that *Escherichia coli* fall into distinct phylogroups.
519 Previously, four phylogenetic groups (A, B1, B2 and D) were considered the main
520 groups of *Escherichia coli* [49], however due to the increased number of *Escherichia*

521 *coli* genomes the phylogenetic groups have been redefined and now consist of seven
522 phylogenetic groups (A, B1, B2, C, D, E, F and G as well as a number of cryptic clades
523 [49,76,77] .The most recognized member of group E is enterohemorrhagic *Escherichia*
524 *coli* (EHEC) including the O157:H7 *Escherichia coli* [78]. The fact that ETEC can
525 belong to group E contradicts the notion that a chromosomal backbone of A, B1 and C
526 groups is necessary for the arrival and maintenance of the genes encoding the two toxins
527 [79] and highlights the fact that ETEC is a highly diverse pathogen.

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

534 Blast analyses confirm that the plasmids identified in this study are often highly
535 homologous to other plasmids identified by either long-read or Illumina sequencing
536 present in GenBank. For instance, the 94.5 kb plasmid pAvM_1441_18 was 98%
537 identical to two 96 kb and 82 kb plasmids belonging to ETEC O25: H16 isolates ATCC
538 43886/E2539C1 and 2014EL-1346-6 sequenced by PacBio by Smith *et al.,* [20] (Figure
539 1e). Plasmid pAvM_E1441_18 is the major virulence plasmid of this lineage carrying
540 genes encoding LT and CS6.

541 The larger plasmid in E1441 (pAvM_E1441_p17) carries both the genes for ETEC CF
542 CS21 and antibiotic resistance determinants. Furthermore, a complete conjugation
543 machinery was present suggesting that this is most likely a mobile plasmid. Movement

544 of such a plasmid results in not only the spread of ETEC virulence genes but also the
545 spread of AMR determinants.

546 Interestingly, Wachsmuth *et al*., [60] analysed transfer frequencies in ETEC O25:H16
547 isolates and found evidence that resistance to tetracycline and sulfathiazole was
548 transferred but not the genes encoding LT [60]. The same study found evidence of two
549 large plasmid of similar size [60] corroborating the finding of two plasmids of similar
550 size in E1441, one with *eltAB* and *cssABCD* without the *tra*-operon and the other with
551 e.g. *sul1* and *tetA* genes on a putative mobile plasmid. Since ATCC 43886/E2539C1,
552 E1441 and 2014EL-1346-6, have been isolated in the 1970's, 1997, and 2014,
553 respectively, this study confirms our findings and indicate that E1441 represent an
554 ETEC lineage with stable plasmid content and putative ability to transfer antibiotic
555 resistance and the CS21 operon by mobilization of one of the plasmids. If this plasmid
556 type is found in other *Escherichia coli* remains to be investigated but the finding that
557 the L4 lineage retains both plasmids in isolates collected over time and world-wide
558 indicate a strong selective force to keep the extrachromosomal contents of both
559 plasmids.

560 The ETEC O169:H41 isolate F8111-1SC3 plasmid unnamed 2 [20,70] is highly similar
561 to pAvM_E1373_28 (Figure 1h). The F8111-ISC3 isolate is part of a CDC collection
562 of ETEC isolates from cruise ship outbreaks and diarrheal cases in US 1996-2003. The
563 antibiotic resistance profiles of these isolates were determined [70] and most isolates
564 of O group 169 were tetracycline resistant corroborating the findings of the *tet* gene in
565 E1373 isolated in Indonesia 1996. ETEC diarrhoea caused by isolates determined to be
566 O169:H41 and STp CS6 positive are repeatedly reported to cause diarrhoea,
567 particularly in Latin America [61,80–82]. Among the cruise ship isolates is the
568 sequenced and characterised virulence plasmid pEntYN10 encoding STp and CS6, that

569 has been described to be unstable and easily lost *in vitro* [80,83]. The E1373 plasmid;
570 AvM_E1373_28 is highly homologous to pEntYN10 (Figure 1h) and the virulence
571 profile of ETEC O169: H41 is conserved in isolates collected globally. Hence, the
572 instability of the plasmid is not biologically valid and current data indicate that plasmids
573 are stable within this linage and serotype.

574 An examination of the cargo carried on prophage and cryptic prophage of our
575 sequenced isolates indicate a number of broad types of proteins present (Table S5).
576 Many have been shown previously to have roles in inhibiting cell division and have a
577 phenotypic signature of antibiotic resistance [72]. It is possible that arresting cell
578 division in the presence of antibiotics preserves the integrity of the bacterial cell until
579 the antibiotic is diluted out over time or degraded.
580 Some, such as *yfdR,* do so by competitive binding to DnaA, a protein involved in
581 chromosomal replication. We found many copies of the *ompX*/*ail* gene in the
582 prophages. Its role is as an outer membrane protein with a role in adhesion to eukaryotic
583 cells and antibiotic resistance make it a desirable gene for phage cargo.
584

585 Two distinctive extra-chromosomal elements which are highly similar to P1 and SSU5
586 phage were identified among the 8 ETEC reference strains sequenced (Figure 2, Figure
587 S3 and Table S5). The SSU5-like element carries a number of genes that allow it to be
588 functional as a plasmid and in fact belongs to the pHCM2-like family of Phage-
589 Plasmids (Figure 2b) [64]. These plasmids are devoid of virulence factors, transposons
590 and antibiotic markers but, they do contain a significant number of DNA metabolism
591 and biosynthesis genes that could be advantageous to the host cell and they may contain
592 bacteriophage inhibitory genes that have not yet been identified. Interestingly, several
593 SSU5 phage-like plasmids have been shown to carry the ESBL gene *bla*CTX-M15 in

594 a number of extra-intestinal pathogenic *Escherichia coli* isolates [66]. ESBL resistance
595 seems to be absent or low in ETEC and the SSU5 phage-like plasmid pAvM_E1373_29
596 does not contain antibiotic resistance genes.
597
598 Classification of the plasmids was made using two approaches, the PlasmidFinder and
599 pMLST. For some of the plasmids a single replicon was found and for others multiple
600 replicons were identified. Further analyses of the multi-replicon plasmids were
601 undertaken combining the classification results with manual curation. For plasmids
602 with different replicons it is evident that these indeed are multi-replicon plasmids
603 (Table 2). Multi-replicon plasmids have been described as a way to broaden their host
604 range, *i.e.* possibility to be transferred between bacteria of different phylogenetic
605 groups. However, plasmids with two replicons of the same Inc group is an indication
606 that these are rather co-integrates or hybrid plasmids. Co-integrated plasmids are a
607 result of a fusion of two plasmids with different or same replicon through
608 recombination [84]. The co-integrated plasmid can separate when needed by the
609 bacteria. The latter indicating that there may be plasmids which are co-integrates
610 similar to the pCoo (CS1 containing plasmid). The pCoo was found to have two
611 replicons, IncFII and IncI1, which likely was formed by recombination of two
612 independent replicons [84]. The plasmids pAvM_E925_6 and pAvM_E925_7 are IncI1
613 and IncFII plasmids, respectively. A blastn comparison with pCoo show that these may
614 be co-integrates (Figure S6). There is a region that has been duplicated in
615 pAvM_E925_6 which is absent in pCoo (Figure S6).

616 **Conclusion**

617 We provide fully assembled, manually cured and annotated chromosomes and plasmids
618 that will serve as new ETEC reference genomes. The in-depth analysis of gene content

619 and order and correct annotations of plasmids will also help to elucidate other plasmids
620 with and without virulence factors in related bacterial species. Our genomes were
621 compared to other long-read sequenced ETEC genomes and we can confirm that the
622 major ETEC lineages all harbor conserved plasmids that have been associated to their
623 respective background genomes for decades. This confirms that the plasmids and
624 chromosomes of ETEC are both crucial for ETEC virulence and superiority as
625 pathogens.
626
627 **Supplementary information**
628 Figure S1-S6
629 Tables S1-S5
630 Additional file 1: Detailed description of ETEC reference plasmids
631
632 **Acknowledgements**
633 No acknowledgements to mention.
634
635 **Authors' contributions**
636 AvM conceived and designed the experiments, performed the experiments, analysed
637 the data, contributed reagents/materials/analysis tools, prepared figures and/or tables,
638 authored the paper and approved the final draft.
639 GB and AvM annotated all IS elements, transposons as well as other mobile elements,
640 contributed to the paper and approved the final draft.
641 DP and AvM annotated the identified prophages, contributed to the paper and approved
642 the final draft.

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

654 **Funding**

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

663 **Availability of data and materials**

664 The datasets supporting the conclusions of this article is included within the articles
665 and its Supplementary information. The sequencing data generated in this study has
666 been submitted to EMBL, accessions pending. The physical ETEC reference strains

667 can be requested by contacting the corresponding author Astrid von Mentzer
668 (avm@sanger.ac.uk or mentzerv@chalmers.se).
669
670 **Ethics approval and consent to participate**
671 Not applicable.
672
673 **Consent for publication**
674 Not applicable.
675
676 **Competing interests**
677 The authors declare that there are no conflicts of interest.
678
679 **Author details**
680 [1]Department of Microbiology and Immunology, Sahlgrenska Academy, University of
681 Gothenburg, Medicinaregatan 7, 413 90 Gothenburg, Sweden
682 [2]Wellcome Sanger Institute, Hinxton, CB10 1SA, Cambridge, UK
683 [3]EMBL-EBI, Hinxton, CB10 1SA, Cambridge, UK
684 [4] University of Cambridge Department of Medicine, Addenbrookes Hospital
685 Box 157, Hills Rd, CB2 0QQ, Cambridge, UK
686 [5] Department of Microbiology, Tumor and Cell Biology, Biomedicum, Karolinska
687 Institutet, Solnavägen 9, 171 65 Solna, Stockholm, Sweden
688 [6]Quadram Institute Bioscience, Norwich Research Park, NR4 7UQ Norwich, UK
689 [#]Currently affiliated with Chalmers University of Technology, Chalmers Tvärgata 3,
690 413 58 Gothenburg, Sweden
691

692 **References**

693

694 1. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity
695 and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the
696 Global Burden of Disease Study 1990-2016. The Lancet Infectious diseases.
697 2018;18:1229–40.

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

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

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

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

708 6. Harro C, Bourgeois AL, Sack D, Walker R, DeNearing B, Brubaker J, et al. Live
709 attenuated enterotoxigenic Escherichia coli (ETEC) vaccine with dmLT adjuvant
710 protects human volunteers against virulent experimental ETEC challenge. Vaccine.
711 2019;37:1978–86.

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

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

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

721 10. Nada RA, Shaheen HI, Khalil SB, Mansour A, El-Sayed N, Touni I, et al.
722 Discovery and phylogenetic analysis of novel members of class b enterotoxigenic
723 Escherichia coli adhesive fimbriae. Journal of clinical microbiology. 2011;49:1403–
724 10.

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

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

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

735 14. Pichel M, Binsztein N, Viboud G. CS22, a novel human enterotoxigenic
736 Escherichia coli adhesin, is related to CS15. Infection and immunity. 2000;68:3280–
737 5.

738 15. Joffre E, von Mentzer A, Ghany MAE, Oezguen N, Savidge T, Dougan G, et al.
739 Allele variants of enterotoxigenic Escherichia coli heat-labile toxin are globally
740 transmitted and associated with colonization factors. Journal of bacteriology.
741 2015;197:392–403.

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

746 17. Joffre E, von Mentzer A, Svennerholm A-M, Sjöling A. Identification of new
747 heat-stable (STa) enterotoxin allele variants produced by human enterotoxigenic
748 Escherichia coli (ETEC). International journal of medical microbiology : IJMM.
749 2016;306:586–94.

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

753 19. Crossman LC, Chaudhuri RR, Beatson SA, Wells TJ, Desvaux M, Cunningham
754 AF, et al. A commensal gone bad: complete genome sequence of the prototypical
755 enterotoxigenic Escherichia coli strain H10407. Journal of bacteriology.
756 2010;192:5822–31.

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

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

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

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

770 24. Sahl JW, Steinsland H, Redman JC, Angiuoli SV, Nataro JP, Sommerfelt H, et al.
771 A comparative genomic analysis of diverse clonal types of enterotoxigenic
772 Escherichia coli reveals pathovar-specific conservation. Infection and immunity.
773 2011;79:950–60.

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

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

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

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

786 29. Sjöling Å, Wiklund G, Savarino SJ, Cohen DI, Svennerholm A-M. Comparative
787 analyses of phenotypic and genotypic methods for detection of enterotoxigenic
788 Escherichia coli toxins and colonization factors. Journal of clinical microbiology.
789 2007;45:3295–301.

790 30. EUCAST EC for AST. Determination of minimum inhibitory concentrations
791 (MICs) of antibacterial agents by broth dilution. 2003 p. ix–xv.

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

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

798 33. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
799 assemblies from short and long sequencing reads. PLoS computational biology.
800 2017;13:e1005595.

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

803 35. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics.
804 2014;30:2068–9.

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

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

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

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

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

819 41. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD
820 2017: expansion and model-centric curation of the comprehensive antibiotic
821 resistance database. Nucleic acids research. 2017;45:D566–73.

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

825 43. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al.
826 Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother.
827 2012;67:2640–4.

828 44. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better,
829 faster version of the PHAST phage search tool. Nucleic acids research. 2016;44:W16-
830 21.

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

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

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

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

841 49. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the
842 Escherichia coli phylogenetic group. Applied and environmental microbiology.
843 2000;66:4555–8.

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

847 51. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The
848 comprehensive antibiotic resistance database. Antimicrobial agents and
849 chemotherapy. 2013;57:3348–57.

850 52. Kumar P, Luo Q, Vickers TJ, Sheikh A, Lewis WG, Fleckenstein JM. EatA, an
851 Immunogenic Protective Antigen of Enterotoxigenic Escherichia coli, Degrades
852 Intestinal Mucin. Infection and immunity. 2014;82:500–8.

853 53. Patel SK, Dotson J, Allen KP, Fleckenstein JM. Identification and molecular
854 characterization of EatA, an autotransporter protein of enterotoxigenic Escherichia
855 coli. Infection and immunity. 2004;72:1786–94.

856 54. Fleckenstein JM, Roy K, Fischer JF, Burkitt M. Identification of a two-partner
857 secretion locus of enterotoxigenic Escherichia coli. Infection and immunity.
858 2006;74:2245–58.

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

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

866 57. Pilonieta MC, Bodero MD, Munson GP. CfaD-dependent expression of a novel
867 extracytoplasmic protein from enterotoxigenic Escherichia coli. Journal of
868 bacteriology. 2007;189:5060–7.

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

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

875 60. Wachsmuth K, Wells J, Shipley P, Ryder R. Heat-labile enterotoxin production in
876 isolates from a shipboard outbreak of human diarrheal illness. Infect Immun.
877 1979;24:793–7.

878 61. Hazen TH, Nagaraj S, Sen S, Permala-Booth J, Canto FD, Vidal R, et al. Genome
879 and Functional Characterization of Colonization Factor Antigen I- and CS6-Encoding
880 Heat-Stable Enterotoxin-Only Enterotoxigenic Escherichia coli Reveals Lineage and
881 Geographic Variation. mSystems. 2019;4:209.

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

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

888 64. Octavia S, Sara J, letters RLF microbiology, 2015. Characterization of a large
889 novel phage-like plasmid in Salmonella enterica serovar Typhimurium.
890 academic.oup.com. n.d.;

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

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

897 67. Colavecchio A, Jeukens J, Freschi L, Rheault J-GE, Kukavica-Ibrulj I, Levesque
898 RC, et al. Complete Genome Sequences of Two Phage-Like Plasmids Carrying the

899 CTX-M-15 Extended-Spectrum β-Lactamase Gene. Genome announcements.
900 2017;5:90.

901 68. Jacoby GA. Mechanisms of Resistance to Quinolones. Clin Infect Dis.
902 2005;41:S120–6.

903 69. Gilcrease EB, Casjens SR. The genome sequence of Escherichia coli tailed phage
904 D6 and the diversity of Enterobacteriales circular plasmid prophages. Virology.
905 2018;515:203–14.

906 70. Beatty ME, Bopp CA, Wells JG, Greene KD, Puhr ND, Mintz ED. Enterotoxin-
907 producing Escherichia coliO169:H41, United States. Emerging infectious diseases.
908 2004;10:518–21.

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

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

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

918 74. Gómez-Duarte OG, Kaper JB. A plasmid-encoded regulatory region activates
919 chromosomal eaeA expression in enteropathogenic Escherichia coli. Infect Immun.
920 1995;63:1767–76.

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

923 76. Clermont O, Gordon DM, Brisse S, Walk ST, Denamur E. Characterization of the
924 cryptic Escherichia lineages: rapid identification and prevalence. Environmental
925 microbiology. 2011;13:2468–77.

926 77. Clermont O, Dixit OVA, Vangchhia B, Condamine B, Dion S, Bridier-Nahmias
927 A, et al. Characterization and rapid identification of phylogroup G in Escherichia coli
928 , a lineage with high virulence and antibiotic resistance potential. Environ Microbiol.
929 2019;21:3107–17.

930 78. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, et al.
931 Comparative genomics reveal the mechanism of the parallel evolution of O157 and
932 non-O157 enterohemorrhagic Escherichia coli. Proceedings of the National Academy
933 of Sciences of the United States of America. 2009;106:17939–44.

934 79. Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Bouguenec CL,
935 Denamur E. A specific genetic background is required for acquisition and expression
936 of virulence factors in Escherichia coli. Molecular biology and evolution.
937 2004;21:1085–94.

938 80. Nishikawa Y, Helander A, OGASAWARA J, MOYER NP, HANAOKA M,
939 HASE A, et al. Epidemiology and properties of heat-stable enterotoxin-producing
940 Escherichia coliserotype O169[ratio]H41. Epidemiology and infection. 1998;121:31–
941 42.

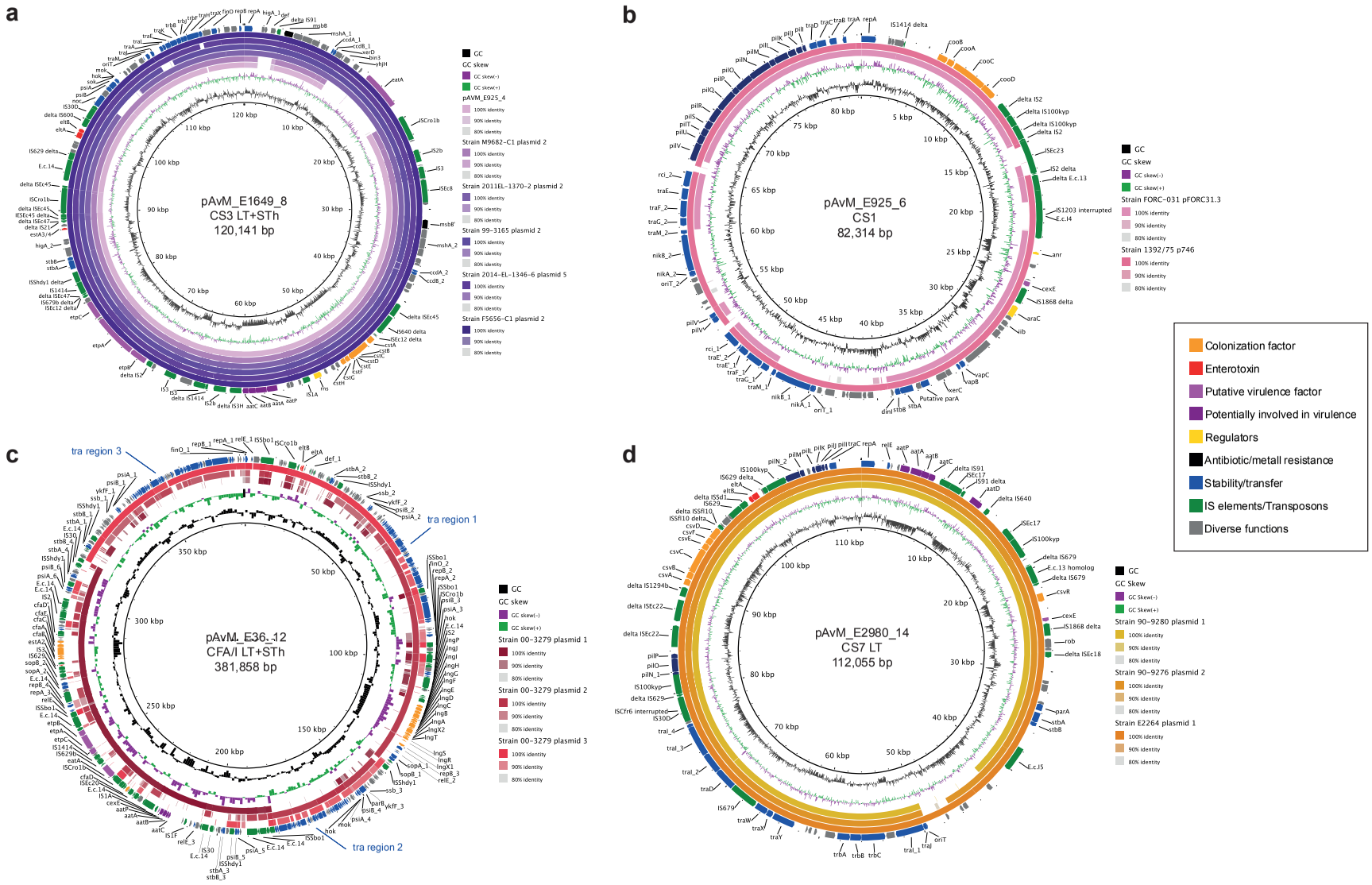
942 81. Torres OR, González W, Lemus O, Pratdesaba RA, Matute JA, Wiklund G, et al.
943 Toxins and virulence factors of enterotoxigenic Escherichia coli associated with
944 strains isolated from indigenous children and international visitors to a rural
945 community in Guatemala. Epidemiol Infect. 2014;143:1662–71.

946 82. Sack DA, Shimko J, Torres O, Bourgeois AL, Francia DS, Gustafsson B, et al.
947 Randomised, double-blind, safety and efficacy of a killed oral vaccine for
948 enterotoxigenic E. Coli diarrhoea of travellers to Guatemala and Mexico. Vaccine.
949 2007;25:4392–400.

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

953 84. Froehlich B, Parkhill J, Sanders M, Quail MA, Scott JR. The pCoo plasmid of
954 enterotoxigenic Escherichia coli is a mosaic cointegrate. Journal of bacteriology.
955 2005;187:6509–16.

956


a
pAvM_E1649_8
CS3 LT+STh
120,141 bp
GC
GC skew
GC skew(-)
GC skew(+)
pAVM_E925_4
Strain M9682-C1 plasmid 2
Strain 2011EL-1370-2 plasmid 2
Strain 99-3165 plasmid 2
Strain 2014-EL-1346-6 plasmid 5
Strain F5656-C1 plasmid 2
100% identity
90% identity
80% identity
b
pAvM_E925_6
CS1
82,314 bp
Strain FORC-031 pFORC31.3
Strain 1392/75 p746
Colonization factor
Enterotoxin
Putative virulence factor
Potentially involved in virulence
Regulators
Antibiotic/metall resistance
Stability/transfer
IS elements/Transposons
Diverse functions
c
tra region 3
tra region 1
tra region 2
pAvM_E36_12
CFA/I LT+STh
381,858 bp
Strain 00-3279 plasmid 1
Strain 00-3279 plasmid 2
Strain 00-3279 plasmid 3
d
pAvM_E2980_14
CS7 LT
112,055 bp
Strain 90-9280 plasmid 1
Strain 90-9276 plasmid 2
Strain E2264 plasmid 1

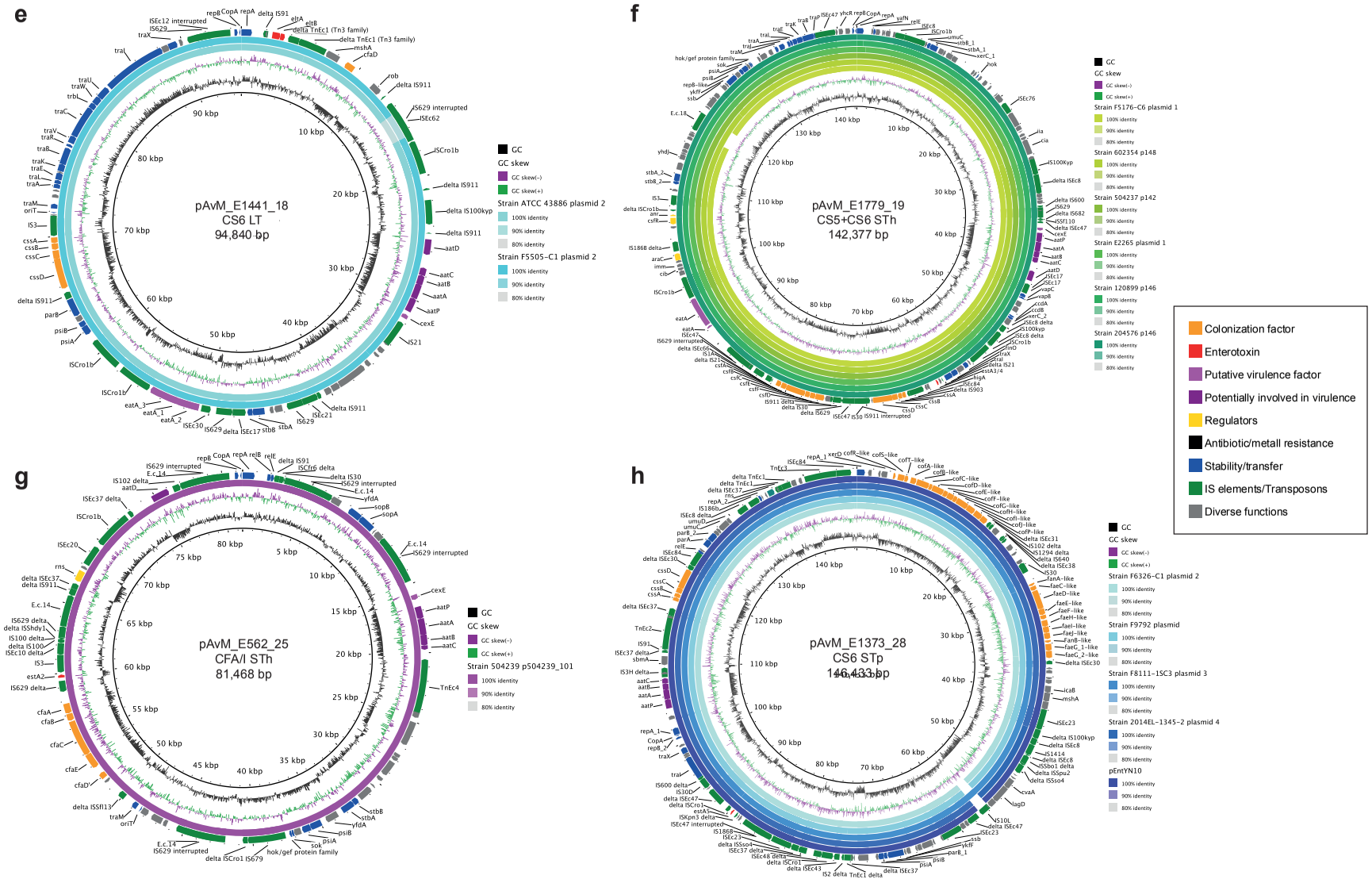
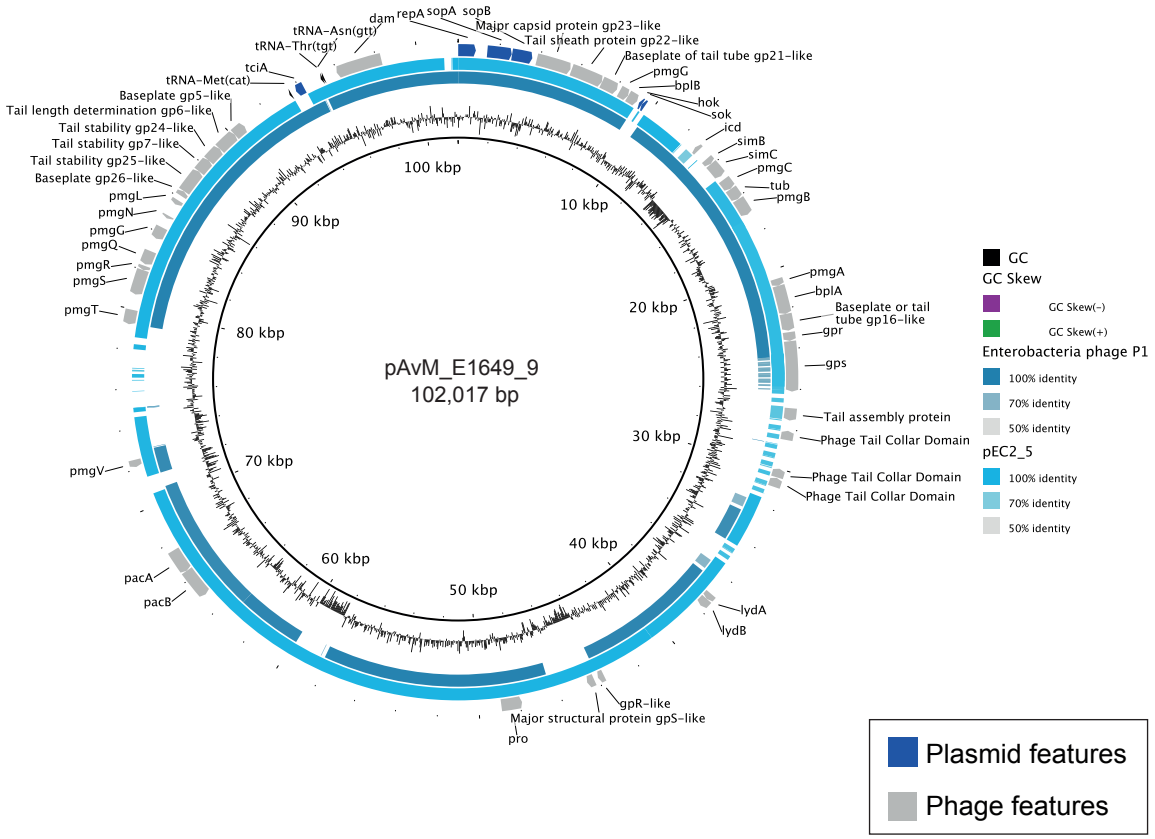

e
pAvM_E1441_18
CS6 LT
94,840 bp
GC
GC skew
GC skew(−)
GC skew(+)
Strain ATCC 43886 plasmid 2
100% identity
90% identity
80% identity
Strain F5505-C1 plasmid 2
100% identity
90% identity
80% identity
f
pAvM_E1779_19
CS5+CS6 STh
142,377 bp
GC
GC skew
GC skew(−)
GC skew(+)
Strain F5176-C6 plasmid 1
100% identity
90% identity
80% identity
Strain 602354 p148
100% identity
90% identity
80% identity
Strain 504237 p142
100% identity
90% identity
80% identity
Strain E2265 plasmid 1
100% identity
90% identity
80% identity
Strain 120899 p146
100% identity
90% identity
80% identity
Strain 204576 p146
100% identity
90% identity
80% identity
Colonization factor
Enterotoxin
Putative virulence factor
Potentially involved in virulence
Regulators
Antibiotic/metall resistance
Stability/transfer
IS elements/Transposons
Diverse functions
g
pAvM_E562_25
CFA/I STh
81,468 bp
GC
GC skew
GC skew(−)
GC skew(+)
Strain 504239 p504239_101
100% identity
90% identity
80% identity
h
pAvM_E1373_28
CS6 STp
146,433 bp
GC
GC skew
GC skew(−)
GC skew(+)
Strain F6326-C1 plasmid 2
100% identity
90% identity
80% identity
Strain F9792 plasmid
100% identity
90% identity
80% identity
Strain F8111-15C3 plasmid 3
100% identity
90% identity
80% identity
Strain 2014EL-1345-2 plasmid 4
100% identity
90% identity
80% identity
pEntYN10
100% identity
90% identity
80% identity

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: M9682-C1 plasmid 2 (1975, USA; CP024277.1), 2011EL-1370-2 plasmid 2 (2011, USA; CP022914.1), 99-3165 plasmid 2 (USA; CP029980.1), 2014-EL-1346-6 plasmid 5 (2014, USA; CP024237.1) and F5656-C1 plasmid 2 (USA, CP024262.1). **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 ATCC 43886 plasmid 2 (CP024255.1) and F5505-C1 plasmid 2 (2013, Sweden; CP023259.1). **f)** pAvM_E1779_19 (CS5+CS6) compared to F5176-C6 plasmid 1 (1997; CP024668.1), 602354 p148 (2009, Bangladesh; CP025848.1), 504237 p142 (2010, India; CP025863.1), E2265 plasmid 1 (2006, Bangladesh; CP023347.1), 120899 p146 (2012, Gambia; CP025917.1), and 204576 p146 (2010, Mali; CP025908.1). **g)** pAvM_E562_25 (CFA/I) compared to p504239_101 (2010, India; CP025860.1). **h)** pAvM_E1373_28 (CS6) compared to F6326-C1 plasmid 2 (1998, USA; CP024265.1), F9792 plasmid (USA; CP023274.1), F8111-1SC3 plasmid 3 (USA; CP024272.1), 2014EL-1345-2 plasmid 4 (2014, USA; CP024227.1) and pEntYN10 (1991, Japan; AP014654.2). The thresholds chosen for the blastn are shown in the key to the right of each plasmid comparison. The colour code for the annotations are listed at the bottom of the figure.

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
dam
repA
sopA
sopB
Majpr 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)
GC
GC Skew
GC Skew(-)
GC Skew(+)
Enterobacteria phage P1
100% identity
70% identity
50% identity
pEC2_5
100% identity
70% identity
50% identity
Plasmid features
Phage features

b

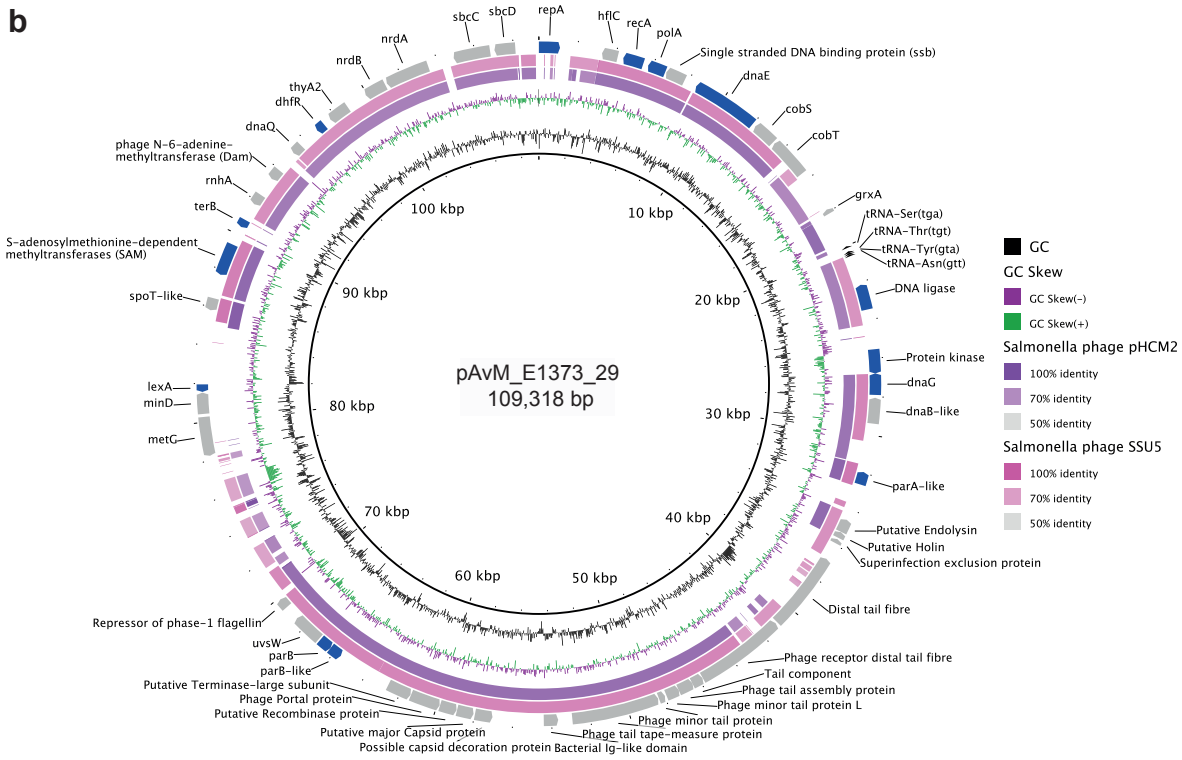

pAvM_E1373_29
109,318 bp
100 kbp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
sbcC
sbcD
repA
hflC
recA
polA
Single stranded DNA binding protein (ssb)
dnaE
cobS
cobT
grxA
tRNA-Ser(tga)
tRNA-Thr(tgt)
tRNA-Tyr(gta)
tRNA-Asn(gtt)
DNA ligase
Protein kinase
dnaG
dnaB-like
parA-like
Putative 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
metC
minD
lexA
spoT-like
S-adenosylmethionine-dependent methyltransferases (SAM)
terB
rnhA
phage N-6-adenine-methyltransferase (Dam)
dnaQ
dhfR
thyA2
nrdB
nrdA
GC
GC Skew
GC Skew(-)
GC Skew(+)
Salmonella phage pHCM2
100% identity
70% identity
50% identity
Salmonella phage SSU5
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 (*Escherichia 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 BRIGS with the thresholds indicated to the right of each plasmid comparison. Selected phage and plasmid annotations are shown in the outer ring.

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: Diarrhea 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[b]
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*	-	
	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)	-	-	
	pAvM_E925_6	82 314	47.8	I1	*iib* (colicin 1b), *repA, stbAB, tra* genes, *pil* locus, *vapBC*	*cooB, A, C, D* (CS1)	*cexE*	-	
	pAvM_E925_7	51 418	45.4	FII	*ccdAB, psiA, repAB, stbAB, tra* genes	-	*eatA_1-5* (two disrupted *eatA* copies)	-	
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*	-	
	pAvM_E1649_9*	102 017	47.6	Y	P1 addiction system (phage related), *repA, sopAB*		-	-	
	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)	-	-	
	pAvM_E1649_11 *	8 834	42.9	No hits	ND	-	-	-	
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*	-	
	pAvM_E36_13	99 448	51.6	B/O/K/Z	*stbAB, relE, repA, pil* genes, *psiAB*	-	-	*tetA, B, C, R mdf(A)-like*	
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*	-	
	pAvM_E2980_15	72 255	52.4	FII	*psiAB, relE, repAB, stbAB, tra* genes	-	-	*strA, strB, sul2, blaTEM-1B*	
	pAvM_E2980_16	48 305	50.3	I1-like	*stbAB, repA, vapBC* (TA-system)	-	*eatA, etpBAC*	-	
L4 **E1441**	pAvM_E1441_17	130 302	51.3	FII + FIB	*pemI/K* (TA-system), *psiAB, repA* (2), *repB, sopAB, srnAC*	*lngX1, R, S, T, X2, A-J, P* (CS21)	-	*aadA1, tetR, tetA, sul1, dfrh1*	

CS6+CS21 LTh (LT17)					(antisense RNA-regulated system), *stbAB, tra* genes				
	pAvM_E1441_18	94 840	47.1	FII	*parB, psiAB, repAB, stbAB*	*cssA, B, C, D* (CS6), *eltAB17* (LTh)	*etaA_1, eatA_2, cexE*	-	
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*	-	
	pAvM_E1779_20	88 759	51.8	FII	*hok-sok* (antisense RNA-regulated system), *psiAB, tra* genes, *repAB, stbAB*	*eltAB15* (LTh)	-	-	
	pAvM_E1779_21	82 464	51.0	FIIY	*repA* (2), *repB, tra* genes, *psiAB, parB, sopAB*	-	-	-	
	pAvM_E1779_22	61 528	50.5	FII	*repAB, tra* genes	-	-		
L6 E562 CFA/I+CS21 STh	pAvM_E562_23	109 853	50.5	I1 (+ FII)	*parB, psiAB, stbAB, pil* genes, *tra* genes	-	*eatA, etpBAC*	-	-
	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)	-	-	
	pAvM_E562_25*	95 616	46.7	FII	*psiAB* (truncated *psiA*) *relE/B* (toxin-antitoxin system), *repAB, stbAB, sopAB*	*cfaA, B, C, D* (CFA/I), *estA2* (STh)	-	-	
	pAvM_E562_26	88 318	52.9	B/O/K/Z	*pil* genes, *pndAC* (antisense RNA-regulated system) *psiAB, relE, repA, tra* genes	-	-	-	
	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)	
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	-	-	
	pAvM_E1373_29	109 318	46.4	FIB	*parB, parB-like, repA*	-	-	-	

[a]Incompatibility groups were determined by PlasmidFinder [35] and pMLST [35]for subtyping.
[b]GCA_000000000 (submitted to EMBL, accessions pending)
*Plasmids not circularised