TITLE

Genome-wide analysis in *Escherichia coli* unravels an unprecedented level of genetic homoplasy associated with cefotaxime resistance.

Jordy P.M. Coolen[1#], Evert P.M. den Drijver[2,3#], Jaco J. Verweij[3], Jodie A. Schildkraut[1], Kornelia Neveling[4], Willem J.G. Melchers[1], Eva Kolwijck[1], Heiman F.L. Wertheim[1^], Jan A.J.W. Kluytmans[2,5,6^], Martijn A. Huynen[7]

[1] Department of Medical Microbiology and Radboudumc Center for Infectious Diseases, Radboud University Medical Center, Nijmegen, The Netherlands.

[2] Department of Infection Control, Amphia Ziekenhuis, Breda, The Netherlands.

[3] Laboratory for Medical Microbiology and Immunology, Elisabeth-Tweesteden Hospital, Tilburg, The Netherlands.

[4]Department of Human Genetics, Radboud University Medical Center, Nijmegen, The Netherlands.

[5] Laboratory for Microbiology, Microvida, Location Breda, The Netherlands.

[6] Julius Center for Health Sciences and Primary Care, UMCU, Utrecht, The Netherlands.

[7] Centre for Molecular and Biomolecular Informatics, Radboud University Medical Center, Nijmegen, The Netherlands.

[#]These authors share first authorship

^These authors contributed equally

ABSTRACT

Cefotaxime (CTX) is a commonly used third-generation cephalosporin (3GC) to treat infections caused by *Escherichia coli*. Two genetic mechanisms have been associated with 3GC resistance in *E. coli*. The first is the conjugative transfer of a plasmid harboring antibiotic resistance genes. The second is the introduction of mutations in the promoter region of the *ampC* β-lactamase gene that cause chromosomal-encoded β-lactamase hyperproduction. A wide variety of promoter mutations related to AmpC hyperproduction have been described. However, their link to a specific 3GC such as CTX resistance has not been reported. Here, we measured CTX MICs in 172 cefoxitin resistant *E. coli* isolates and performed genome-wide analysis of homoplastic mutations associated with CTX resistance by comparing Illumina whole-genome sequencing data of all isolates to a PacBio tailored-made reference chromosome. We mapped the mutations on the reference chromosome and determined their occurrence in the phylogeny, revealing extreme homoplasy at the -42 position of the *ampC* promoter. The 24 occurrences of a "T" at the -42 position rather than the wild type "C", resulted from 18 independent C>T mutations in 5 phylogroups. The -42 C>T mutation was only observed in *E. coli* lacking a plasmid-encoded *ampC* gene. The association of the -42 C>T mutation with CTX resistance was confirmed to be significant (FDR < 0.05). To conclude, genome-wide analysis of homoplasy in combination with CTX resistance identifies the -42 C>T mutation of the *ampC* promotor as significantly associated with CTX resistance and underline the role of recurrent mutations in the spread of antibiotics resistance.

Keywords

Escherichia Coli, Genomics, Whole genome sequencing, *ampC,* Bioinformatics

Abbreviations

3GC, third-generation cephalosporin; *campC*, chromosomal-mediated ampC; CAT, computerized adaptive testing; CTX, cefotaxime; DNA, Deoxyribonucleic acid;EHEC, enterohemorrhagic *Escherichia Coli*; ESBL, extended-spectrum β-lactamases; FOX, cefoxitin; gDNA, genomic DNA; MICs, minimal inhibitory concentrations;MLST, multilocus sequence typing; *pampC*, plasmid-mediated ampC; PG, peptidoglycan; qRT-PCR, quantitative reverse transciptase polymerase chain reaction; SMRT, Single-molecule real-time sequencing; SNP, single-nucleotide polymorphism; ST, sequence type; UPEC, uropathogene *Escherichia Coli*; WGS, whole genome sequencing

Impact Statement

In the past decades, the worldwide spread of extended spectrum beta-lactamases (ESBLs) has led to a substantial increase in the prevalence of resistant common pathogens, thereby restricting available treatment options. Although acquired resistance genes, e.g. ESBLs, get most attention, chromosome-encoded resistance mechanisms may play an important role as well. In E. coli chromosome-encoded β-lactam resistance can be caused by alterations in the promoter region of the *ampC* gene. To improve our understanding of how frequently these alterations occur, a comprehensive interpretation of the evolution of these mutations is essential. This study is the first to apply genome-wide homoplasy analysis to better perceive adaptation of the E. coli genome to antibiotics. Thereby, this study grants insights into how chromosomal-encoded

69 antibiotic resistance evolves and, by combining genome-wide association studies with homoplasy
70 analyses, provides potential strategies for future association studies into the causes of antibiotics
71 resistance.
72
73 **Data summary**
74 All data is available under BioProject: PRJNA592140. Raw Illumina sequencing data and metadata
75 of all 171 *E. coli* isolates used in this study is available from the Sequence Read Archive database
76 under accession no. SAMN15052485 to SAMN15052655. Full reference chromosome of
77 ampC_0069 is available via GenBank accession no. CP046396.1 and NCBI Reference Sequence:
78 NZ_CP046396.1.
79
80 **Introduction**
81 *Escherichia coli* is an important pathogen in both community and healthcare-associated
82 infections [1,2]. In the past decades, a substantial increase in resistance to third-generation
83 cephalosporin (3GC) antibiotics in *E. coli* has been observed worldwide, mainly caused by the
84 production of extended-spectrum β-lactamases (ESBL) and AmpC β-lactamases, restricting
85 available treatment options for common infections [3]. AmpC β-lactamases differ from ESBL as
86 they not only hydrolyze broad-spectrum penicillins and cephalosporins, but also cephamycins.
87 Moreover, AmpC β-lactamases are not inhibited by ESBL-inhibitors like clavulanic acid [3], limiting
88 antibiotic treatment options even further. A widely used screening method for AmpC production
89 is the use of cefoxitin (FOX) susceptibility, a member of the cephamycins [4].
90 Although *ampC* β-lactamase genes can be plasmid-encoded (*pampC*), they are also

91 encoded on the chromosomes of numerous Enterobacterales. *E. coli* naturally carries a
92 chromosomal-mediated *ampC* (*campC*) gene, but unlike most other Enterobacterales this gene
93 is non-inducible due to the absence of the *ampR* regulator gene [3]. Consequently, chromosomal
94 AmpC production in *E. coli* is exclusively regulated by promoter and attenuator mechanisms. This
95 results in constitutive low-level c*ampC* expression that still allows the use of 3GC antibiotics, such
96 as cefotaxime (CTX), to treat *E. coli* infections [3]. However, various mutations in the
97 promoter/attenuator region of *E. coli* may cause constitutive hyperexpression of *campC* [5,6],
98 thereby increasing the Minimal Inhibitory Concentrations (MICs) for broad-spectrum penicillins
99 and cephalosporins and limiting appropriate treatment options.

100 A wide variety of promoter and attenuator mutations have been related to AmpC
101 hyperproduction [6]. AmpC hyperproduction is primarily caused by alterations of the *ampC*
102 promoter region, leading to a promoter sequence that more closely resembles the *E. coli*
103 consensus sigma 70 promoter with a TTGACA -35 box separated by 17 bp from a TATAAT -10 box.
104 These alterations can be divided into different variants associated with e.g. an alternate displaced
105 promoter box, a promotor box mutation or an alternate spacer length due to insertions [6].
106 Furthermore, mutations of the attenuator sequence can lead to changes in the hairpin structure
107 that strengthen the effect of promoter alterations on AmpC hyperproduction. In a study on
108 cefoxitin-resistant *E. coli* isolated from Canadian hospitals, Tracz *et al.* described 52 variants of
109 the promoter and attenuator region [6]. In this study a two-step quantitative reverse
110 transcriptase (qRT-)PCR was used to determine the effect of promoter/attenuator variants on
111 *ampC* expression. Various mutations were related to different delta–delta cycle threshold values
112 in the RT-PCR and corresponding variations in FOX resistance. An interesting observation that

113 emerged from this study was that the -32T>A and the -42C>T mutation were the major
114 alterations that strengthened the *ampC* promoter. Both result in a consensus -35 box. Although
115 it is known that AmpC hyperproduction leads to FOX resistance as studied by Tracz *et al.,* the
116 effect of various mutations on resistance to a 3GC antibiotic such as CTX have not been explored.
117 This is relevant because CTX is commonly used in the treatment of patients with severe *E. coli*
118 infections, often in combination with selective digestive tract decontamination in Intensive Care
119 Units [7,8].

120 While previous research mainly focused on the chromosomal AmpC resistance
121 mechanism and the impact of AmpC hyperproduction, there is a lack in knowledge and
122 understanding of the evolutionary origin of these promoter/attenuator variants. Notably, it is
123 unexplored how the two most prominent promoter mutations -32T>A and -42C>T are distributed
124 over the *E. coli* phylogeny and therewith how often they occur. More precisely, literature shows
125 selective pressure can lead to convergent evolution that results in the reoccurrence of a mutation
126 in multiple isolates independently and in separate lineages [9]. This phenomenon is named
127 homoplasy [10]. A Consistency Index can be calculated to quantify homoplasy by dividing the
128 minimum number of changes on the phylogeny by the number of different nucleotides observed
129 at that site minus one [11], effectively quantifying how often the same mutation occurred in a
130 phylogenetic tree. One can use the Consistency Index to recognize genomic locations subjected
131 to homoplasy, and relate the single-nucleotide polymorphism (SNP) positions that are
132 inconsistent with the phylogeny to antibiotic resistance, as has e.g. been done in multiple studies
133 on *Mycobacteria* spp [12–15].

134 In the present study, we hypothesize that some of the mutations in the *ampC*
135 promoter/attenuator region are homoplastic and are associated with CTX resistance. To test our
136 hypothesis, we performed genome-wide homoplasy analysis and combined it with a genome-
137 wide analysis of polymorphisms associated with CTX resistance by constructing a tailored *E. coli*
138 reference chromosome and combining it with WGS data of 172 FOX resistant *E. coli* isolates
139 previously collected by our research group [16].
140

141 **Methods**

142 **Isolate selection, DNA isolation, library preparation and DNA sequencing**

143 One hundred seventy-two *Escherichia coli* isolates previously used by our study group [16] were
144 selected in the present study (see Table S1 in supplemental material). To summarize the method;
145 DNA isolation was performed as previously described [16], library preparations were performed
146 using Illumina Nextera XT library preparation kit (Illumina, San Diego, CA, USA), and DNA
147 sequencing was performed using an Illumina NextSeq 500 (Illumina, San Diego, CA, USA) to
148 generate 2x 150bp paired-end reads or 2x 300 bp reads on an Illumina MiSeq (Illumina, San Diego,
149 CA, USA). *De novo* assembly was also performed identical to the method as described in Coolen
150 *et al.* 2019 [16] using SPAdes version 3.11.1 [17].
151

152 **Phylogroup and MLST**

153 Phylogroup stratification was performed using ClermonTyping version 1.4.0 [18]. MLST STs were
154 derived from the contigs using mlst version 2.5 pubMLST, 31 October 2017 [19,20].
155

156 **Obtaining the *ampC* promoter/attenuator region**

157 To detect the promoter/attenuator region a custom blast database [21] was created using the
158 271 bp fragment as described by Peter-Getzlaff *et al.*[22] using *Escherichia coli* K-12 strain ER3413
159 (accession: CP009789.1) ABRicate version 0.8.9 [23] was used to locate matching regions per
160 sample and were extracted and converted into multi-fasta format using a custom python script.
161 Strains were labelled AmpC hyperproducer when promoter mutations were found, as reported
162 by Caroff *et al.* [24] and Tracz *et al.* [6].

163

164 **Plasmid-mediated ampC detection**

165 Detection of *pampC* genes was performed by using ABRicate version 0.5 [23] and ResFinder
166 database (2018-02-16) as described by Coolen *et al.* [16].

167

168 **PacBio single molecule real-time (SMRT) sequencing of *E. coli* isolate**

169 For PacBio SMRT sequencing, genomic DNA (gDNA) was extracted using the Bacterial gDNA
170 Isolation Kit (Norgen Biotek Corp., CAN, ON, Thorold). A single *E. coli* isolate was subjected for
171 DNA shearing using Covaris g-TUBEs (Covaris Inc, US, MA, Woburn) for 30 seconds on 11,000
172 RPM (*g*). Each DNA sample was separated into two aliquots. Size selection was performed using
173 a 0.75% agarose cassette and marker S1 on the BluePippin (Sage Science Inc, US, MA, Beverly) to
174 obtain either 4-8 kb or 4-12 kb DNA fragments. This size selection was chosen to maintain all DNA
175 fragments including these originating from plasmids (data not used in this study). Library
176 preparation was performed using the SMRTbell Template prep kit 1.0 (Pacific Biosciences, US,
177 CA, Menlo Park). For cost-effectiveness, samples were barcoded and pooled with other samples

178 that are not relevant for this study. Sequencing was conducted using the PacBio Sequel I (Pacific
179 Biosciences, US, CA, Menlo Park) on a Sequel SMRT Cell 1M v2 (Pacific Biosciences, US, CA, Menlo
180 Park) with a movie time of 10 h (and 186 min pre-extension time). Subreads per sample were
181 obtained by extracting the bam files using SMRT Link version 5.1.0.26412 (Pacific Biosciences, US,
182 CA, Menlo Park).
183
184 **Chromosomal reconstruction using *de novo* hybrid assembly**
185 To obtain a full-length chromosome, Unicycler version 0.4.7 [25] (settings: --mode bold) was
186 used, combining Illumina NextSeq 500 2x 150 bp paired-end reads with PacBio Sequel SMRT
187 subreads. Because unicycler requires fasta reads as input, the subreads in bam format were
188 converted to fasta by using bam2fasta version 1.1.1 from pbbioconda
189 (https://github.com/PacificBiosciences/pbbioconda) prior to *de novo* hybrid assembly. The full
190 circular chromosome was uploaded to NCBI and annotated using the NCBI Prokaryotic Genome
191 Annotation Pipeline (PGAP) version 4.10 [26,27].
192
193 **SNP analysis using *E. coli* reference ampC_0069**
194 Alignment of Illumina reads and SNP calling was performed for all isolates to the reference
195 chromosome of *E. coli* isolate ampC_0069 using Snippy version 4.3.6
196 (https://github.com/tseemann/snippy). A full-length alignment (fullSNP) and a coreSNP
197 alignment containing SNP positions shared among all isolates were generated by using snippy-
198 core version 4.3.6 (https://github.com/tseemann/snippy).
199

200 **Inferring of phylogeny**

201 A phylogenetic tree was inferred by using the coreSNP alignment as input for FastTree(MP)
202 version 2.1.3 SSE3 (settings: -nt -gtr) [28].

203

204 **Detection of Homoplasy**

205 The Consistency Index for all nucleotide positions on the chromosome was calculated using
206 HomoplasyFinder version 0.0.0.9000 [10]. The coreSNP phylogeny was used as true phylogeny
207 and the Consistency Index was calculated using the multiple sequence alignments fullSNP
208 alignment as input.

209

210 **Relate mutations to CTX resistance**

211 To assess if certain mutations were linked to CTX resistance all non-plasmid harboring *ampC E.*
212 *coli* isolates were used. CTX resistance was defined using EUCAST guidelines standards of CTX
213 MIC > 2 mg/L [29]. CTX MIC results were obtained from our previous study [16]. For each
214 nucleotide position on the reference chromosome the number of Resistant and Sensitive isolates
215 were counted and tested for adenine vs all other nucleotides, thymine vs all other nucleotides,
216 cytosine vs all other nucleotides, and guanine vs all other nucleotides creating a contingency
217 table and performing a Fisher Exact Test in R 3.6.1 [30]. To correct for multiple testing, *P* values
218 were adjusted using FDR [31].

219

220 **Selection of genomic positions of interest**

221 By combining previous metrics most relevant genomic positions were selected. Criteria for
222 selection are, FDR <= 0.05 to CTX and Consistency Index of <= 0.05882353. Annotation of
223 mutation positions was obtained by using the genome annotation of reference chromosome
224 (GenBank accession no. CP046396) and applying snpEff (version 4.3t) [32]. The EnteroBase core-
225 genome MLST and whole-genome MLST schemes were used to distinguish core and accessory
226 genes [33].
227
228 **Recombination analysis**
229 Gubbins version 2.4.1. (settings: -f 30) was used to detect recombination regions with coreSNP
230 alignment and tree as input [34].
231
232 **Visualization of data**
233 The interactive Tree of Life web-based tool iTOL version 5.3 was used to visualize the
234 phylogenetic tree [35]. Information about CTX resistance, presence of the *pampC* gene, *campC*
235 hyperproduction as defined, MLST and phylogroup, as well as alignments of promoter and
236 attenuator region were incorporated into visualization. The sequence logo of the promoter and
237 the attenuator alignment were generated using the web-based application Weblogo version 3.7
238 [36] (http://weblogo.threeplusone.com). A chromosome ideogram of the *E. coli* isolate
239 ampC_0069 reference chromosome was visualized using CIRCOS software package version 0.69-
240 8 [37]. Consistency Index scores and significant mutations associated with CTX resistance were
241 plotted in the ideogram. Gubbins results were displayed by using phandango [38].
242

243 **Overview of method**

244 A workflow graph of the methods is visualized in Fig 1 using the web-based application yEd Live

245 version 4.4.2 (https://www.yworks.com/yed-live/).

246

247 **RESULTS**

248 ***E. coli* collection**

249 To study genetic homoplasy events in suspected AmpC producing *Escherichia coli,* FOX MIC > 8

250 ug/ml and ESBL phenotype negative *E. coli* isolates (*n*=172) were selected as previously described

251 by Coolen *et al.* [16] (see Table S1 in supplemental material). The entire collection was subjected

252 to whole-genome sequencing followed by *de novo* assembly of the sequence reads to obtain

253 contigs.

254

255 **MLST and phylogroup variants**

256 To access the genetic diversity of our *E. coli* collection, we identified both multi-locus sequence

257 typing (MLST) and phylogroup variants of each of the 172 *E. coli* isolates. Seventy-five different

258 sequence types (STs) were identified, of which ST 131 (8.1% *n*=14), ST 38 (7.0% *n*=12), and ST 73

259 (7.0% *n*=12) were the most prevalent. The sequence types of 13 isolates are unknown.

260 Phylogroup stratification revealed that the isolates belonged to eight different phylogroups

261 (Table 1). Phylogroup B1, B2, and D were the most prevalent. One isolate belonged to *Escherichia*

262 clade IV (st. no. ampC_0128).

263

264 ***ampC* promoter and attenuator variants**

265 We examined the whole *E. coli* genome. However, we firstly focused on mutations in the *ampC*
266 promoter and attenuator region. Previously described mutations in the *ampC* promoter region
267 that according to Tracz *et al.* lead to "hyperproduction" of AmpC were detected in 59 (34.4%) of
268 the isolates [6]. These isolates were therefore labelled as hyperproducers. Analysis of the
269 promotor area (-42 to -8) resulted in 21 different variants and the wild type (see Table 1). In the
270 attenuator region (+17 to +37), 18 different variants were identified (see Table 1). One isolate
271 (ampC_0128) showed an unusual promoter variant, a four-nucleotide deletion (-45_-42delATCC).
272 Moreover, an insertion (21_22insG) of unknown function was detected in the attenuator (see
273 Table S1 in supplemental material).
274
275 **Plasmid-mediated *ampC* variants**
276 As we aim to associate chromosomal mutations with CTX resistance, differentiation of *pampC*
277 harboring isolates from non-*pampC* harboring isolates was required. Genomic analysis showed
278 that 90 (52.3%) of the isolates harbored a *pampC* gene of which $bla_{\text{CMY-2}}$ was the most prevalent
279 (*n*=78). One isolate harbored two different *pampC* genes ($bla_{\text{CMY-4}}$ and $bla_{\text{DHA-1}}$) (ampC_0119).
280 One isolate contained a $bla_{\text{CTX-M-27}}$ gene combined with a $bla_{\text{CMY-2}}$ gene (ampC_0114) but was ESBL
281 disc test negative (see Table S1 in supplemental material). In 23 (13.4%) of the isolates neither
282 *pampC* nor described mutations related to AmpC hyperproduction were detected and are noted
283 as low-level AmpC producers.
284
285 **Tailored reference chromosome**

286 To be able to reconstruct an accurate phylogeny we selected *E. coli* isolate ampC_0069, one of
287 the strains of the study, to use as reference chromosome for SNP calling. The tailored reference
288 chromosome was constructed through a hybrid assembly of *n*=4,423,109 2x 150 bp Illumina
289 NextSeq 500 paired-end reads together with *n*=218,475 PacBio Sequel SMRT subreads (median
290 5,640 bp). This resulted in a high-quality full circular chromosome of *E. coli* isolate ampC_0069,
291 with a size of 5,056,572 bp. This isolate belongs to ST648 and contains a plasmid-encoded bla_{CMY-}
292 $_{42}$. The full circular chromosome was uploaded to GenBank accession no. CP046396 and was used
293 for further analysis. Genome annotation with the NCBI Prokaryotic Genome Annotation Pipeline
294 (PGAP) identified 4,720 Coding Sequences.
295
296 **SNP calling**
297 To be able to reconstruct the phylogeny and obtain SNP positions, we mapped reads of all isolates
298 to the reference chromosome *E. coli* ampC_0069, (accession no. CP046396) resulting in a
299 coreSNP alignment containing 314,200 variable core SNP positions. For further details per isolate
300 see Table S2 in supplemental material. To validate our SNP calling method we compared the
301 ampC_0069 Illumina NextSeq 500 paired-end reads to the reference chromosome of
302 ampC_0069, resulting in 0 SNPs detected, supporting that the SNP calling data and method
303 produce no false positives.
304
305 **Phylogenetic tree based on coreSNP**
306 The coreSNP alignment was used for further analysis. Figure 2 illustrates the approximately
307 maximum-likelihood phylogenetic tree of all 172 isolates based on the coreSNP alignment. The

308 tree has a robust topology as indicated by computerized adaptive testing (CAT) likelihood
309 calculations, resulting in only three positions with a value ≤60% [28]. When focusing on the *ampC*
310 promoter mutations, they were most prevalent in phylogroups B1, B2, and C, although they were
311 present in all phylogroups except phylogroup E that lacked mutations in either the promoter or
312 attenuator region. Interestingly, two positions previously highlighted by Tracz *et al.,* -42 and -32,
313 are only mutated in the absence of a *pampC* gene, even in isolates with a similar MLST (ST12,
314 ST88, and ST131). The -42C>T mutation, which results in an alternate displaced promoter box
315 and therefore leads to increased resistance [6], is present in 24 isolates in 5 distinct phylogroups
316 and in 17 separate phylogenetic branches, indicating that this mutation is homoplastic.
317 Additionally, the -32T>A mutation in the promoter, previously also associated to resistance [6],
318 is present in 20 isolates in 3 distinct phylogroups and in 14 separate phylogenetic branches. To
319 quantify the level of homoplasy we calculated the Consistency Index.
320
321 **Genomic homoplastic mutations**
322 We calculated the Consistency Index for all positions on the *E. coli* reference chromosome. A low
323 Consistency Index value for a position indicates a high degree of inconsistency with the
324 chromosomal phylogeny and can be calculated by HomoplasyFinder as described in earlier
325 studies [10,39,40]. As can be observed in Figure 3, results clearly indicate that notwithstanding
326 multiple other low scoring Consistency Index positions in the promoter and attenuator, position
327 -42 (4,470,140) and -32 (4,470,150) are the lowest scoring, respectively 0.05882353 and
328 0.07142857 (see also Fig. S1 in the supplemental material). To access how extreme these
329 Consistency Index values are, we calculated the Consistency Index for all positions in the

330 chromosome (see Fig. S2 in the supplemental material). All Consistency Indexes < 1.0 are plotted
331 in the outer ring (ring A) of Figure 4. Results show that only 9,640 out of 5,056,572 positions
332 (0.19%) had a Consistency Index ≤ 0.07142857 (see Figure 4 ring A, cutoff is indicated by black
333 circle). This clearly indicates that positions with low Consistency Indexes are rare, but not unique.
334 Although these 9,640 positions have a low Consistency Index, we do not yet know their relation
335 to CTX resistance.
336
337 **CTX resistance measurements**
338 Cefotaxime MIC measurements from Coolen *et al.* in relation to the genotype of the *E. coli*
339 isolates are shown in supplementary table 1. Eighty-four of ninety (93.3%) *pampC* harboring *E.*
340 *coli* were CTX resistant (MIC >2 mg/L) based on EUCAST clinical breakpoints. Twenty-one of fifty-
341 nine (35.6%) isolates categorized as hyperproducers based on Trazc *et al.* were CTX resistant,
342 primarily isolates with the -42 (*n*=15) or -32 mutation (*n*=2). The *pampC* genes never occurred
343 simultaneously with the -42 or -32 mutations in any of these isolates. One of twenty-three (4.3%)
344 isolates categorized as a low-level AmpC producer (no *pampC* gene and no known ampC
345 promoter mutation) tested CTX resistant and contained an insertion in the spacer region (-16_-
346 15insT), which has not been described by Tracz *et al* [6]. As depicted in Figure 2 the non-*pampC*
347 strains with a phenotype of CTX > 2 mg/L were present in all phylogroups, although CTX resistant
348 isolates with the -42 or -32 mutation were predominantly present in phylogroups B1, B2, and C.
349
350 **Geno- to phenotype**

351 To be able to link *E. coli* chromosomal mutations to CTX resistance we excluded all *E. coli* isolates
352 with a plasmid containing an *ampC* β-lactamase gene. The association of SNPs to CTX resistance
353 phenotype (MIC > 2 mg/L) was tested in the remaining 82 isolates using Fisher's Exact Test. After
354 FDR correction to 0.05, 45,998 significant positions were found (see Figure 4 ring B). Mutation
355 C>T on position -42 of the *ampC* promoter was found to be significantly associated to CTX
356 resistance (FDR = 0.034). However, position -32 A>T was not significantly associated to CTX
357 resistance (FDR = 1).
358
359 **Homoplasy-based association analysis**
360 Combining the outcome of the homoplasy analysis with the significant CTX resistance associated
361 positions results in genomic positions associated to CTX resistance that have evolved multiple
362 times independently. After selecting the lowest scoring Consistency Index positions, ≤
363 0.05882353, 24 relevant genomic positions were identified that had both a low Consistency Index
364 and a significant association with CTX resistance. Most notably, one of these 24 positions is
365 position -42. Only two mutations of those 24 that were located in genes were non-synonymous:
366 a (conservative) missense mutation in the type II secretion system protein L (*gspL*) gene leading
367 to Ser330Thr alteration and a mutation in the hydroxyethylthiazole kinase (*thiM*) gene resulting
368 in a Thr122Ala alteration according to the annotation of *E. coli* strain ampC_0069 (accession no.
369 CP046396). In addition to the non-synonymous mutation found on the *gspL* gene, eight
370 synonymous mutations are also located in genes annotated as being part of the type II secretion
371 system. A complete overview is presented in Table 2.
372

373 **Recombination analysis**

374 To verify if the level of homoplasy could be a result of recombination, we used Genealogies
375 Unbiased By recomBinations In Nucleotide Sequences (Gubbins) algorithm to predict
376 recombination events in our isolate collection [34]. This analysis showed frequent recombination
377 events in our 172 *E. coli* isolates (see Fig. S3 in the supplemental material). Results illustrate that
378 recombination blocks cover the region of the *gspL* and the *thiM* gene and their high homoplasy
379 levels could thus be due to recurrent recombination rather than independent mutations.
380 Nonetheless, position -42 in the *ampC* promoter is not located in a region effected by
381 recombination as shown in Fig. S3. Moreover, when inferring the phylogenetic tree corrected for
382 recombination events as obtained from the Gubbins analysis, the -42C>T mutation actually
383 occurred in 18 independent branches rather than the 17 branches in the uncorrected tree. This
384 supports our previous results that this mutation is homoplastic, and not the results of a recurrent
385 recombination event.

386

387 **Discussion**

388 We present a genome-wide analysis in which homoplastic mutations are associated with
389 antibiotic resistance in *E. coli*. By comparing whole-genome sequencing data of 172 *E. coli* isolates
390 to a tailored reference chromosome we were able to reconstruct the evolution of the genomes
391 and therewith map recurrent events, allowing us to detect homoplasy associated to CTX
392 resistance.

393 Our foremost finding is the significant association of the -42C>T mutation, in the *ampC*
394 promoter, to CTX resistance that evolved independently at least 17 times in 5 distinct

395 phylogroups. The -42C>T mutation has been confirmed in former studies to result in AmpC
396 hyperproduction in *E. coli*. Nelson *et al.* demonstrated an 8 to 18 times increase in activity of
397 AmpC when cloning the promoter upstream a *lac* operon [41]. Vice versa, Caroff *et al.* found a
398 decrease in expression of AmpC when cloning the promoter with a -42T>C mutation in a pKK232-
399 8 reporter plasmid with chloramphenicol acetyltransferase gene [24]. Tracz *et al*. confirmed that
400 the -42C>T mutation has the strongest effect on the *ampC* promoter, resulting in a high
401 expression of the *ampC* gene as detected by RT-qPCR [6]. Despite the fact that the -42C>T
402 mutation has such a strong effect on AmpC production the effect of the mutation on CTX MICs
403 had not been confirmed. Moreover, the contribution of convergent evolution on this position
404 relative to the role of the expansion of a clone with a beneficial mutation at this position has not
405 been determined. That being the case, this study provides evidence that this -42 C>T mutation is
406 not a result of a recombination event and most likely evolved many times independently.
407 Remarkably, we observed that the -42C>T mutation never occurs in the presence of a *pampC*
408 gene (in zero out of twenty-four cases). This was even noticed in isolates with the same MLST,
409 i.e. ST88 -42C>T (*n*=3) and *pampC* (*n*=1), suggesting preferred exclusivity for one of the resistance
410 mechanisms. One study mentioned the co-occurrence of the -42C>T mutation and a *pampC* gene
411 in only one out of thirty-six strains [42]. One could speculate that the exclusivity is a matter of
412 what arrives first, the plasmid or the mutation, after which there is no selective advantage for
413 the second mechanism, or that there is actually a fitness cost to having both the mutation and
414 the plasmid relative to having only the mutation or the plasmid.
415

416 The study performed by Tracz *et al.* showed that position -32T>A on the promotor of
417 *ampC* associates with AmpC hyperproduction that results in elevated MIC levels for FOX [6].
418 Surprisingly, in the current study no significant association of -32T>A with CTX resistance was
419 noticed despite its low Consistency Index. Only two out of twenty isolates with the -32T>A were
420 CTX resistant, four out of twenty showed an intermediate elevated CTX MIC, and fourteen were
421 susceptible for CTX. Although we do not know under which conditions this mutation did arise, it
422 can be speculated that the high level of homoplasy at the -32 position is associated with a
423 different trait, e.g. resistance against another antibiotic.
424
425 Prior studies discovered the importance of mutations in the promoter elements. Random
426 sequences can even evolve expression comparable to the wild-type promoter elements after only
427 a single mutation [43]. Furthermore, these promotor elements evolve to only a few forms
428 indicating convergent evolution [44], as also observed in the present study. All encountered
429 variants seem to result in a sequence that resembles the *E. coli* consensus sigma 70 promoter
430 more than the wild type sequence they are derived from [6].
431
432 Next to the -42C>T promoter mutation we detected twenty-three other positions in our
433 analysis that are associated with CTX resistance and have extreme high levels of homoplasy. Most
434 of these are synonymous mutations, with only two missense mutations (*thiM* and *gspL*) found. It
435 is remarkable that one missense mutation (p.Ser330Thr) is located in *gspL* that encodes for a
436 protein of the type II secretion system. The type II secretion system is used by many gram-
437 negative bacteria to translocate folded proteins from the periplasm, through the outer

438 membrane, into the extracellular milieu [45]. The system is composed of 12–15 different general
439 secretory pathway (Gsp) proteins and is related to virulence of various pathogenic *E. coli,* e.g.
440 EHEC and UPEC [46–48]. It could be that in our selection of mainly clinical samples a certain
441 predilection has occurred towards isolates with particular virulence traits. The *gspL* gene has
442 been described as being part of the accessory genome of *E. coli* [49]. Our study supports this
443 finding as some strains did not harbor this gene. Additionally, we found evidence that
444 recombination events in the type II secretion system could be the underlying cause of the
445 extreme homoplasy levels. Still, it is remarkable that missense mutation p.Ser330Thr in the *gspL*
446 gene correlates with the CTX resistance trait even though it is most likely caused by a
447 recombination event. To the best of our knowledge no relationship between type II secretion
448 system and CTX resistance has been observed before. One could hypothesize that the mutation
449 is a secondary adaptation needed to cope with the elevated AmpC production, as the
450 peptidoglycan (PG) layer is effected by AmpC hyperproduction and the type II secretion system
451 contains proteins that are partly localized in the periplasm [50,51].
452
453 The use of genomic data to detect homoplasy events is not an uncommon scientific
454 technique [52–54]. In *Mycobacterium tuberculosis*, it is a well-known method to identify
455 advantageous mutations, as they are likely to be associated with phenotypes such as drug
456 resistance, heightened transmissibility, or host adaptation [12–15]. A similar approach was taken
457 recently by Benjak *et al.* to screen for highly polymorphic genes and genomic regions of
458 *Mycobacterium leprae* [55]. Homoplasy-based association analysis limits phylogenetic bias by
459 correcting for genetic relatedness of strains with the same phenotype, thereby increasing

460 statistical power to find true associations [14]. Taking this into account, the use of homoplasy-
461 based association analysis seems viable to relate polymorphic sites to phenotypic traits in
462 bacteria. Still, studies on other genera than mycobacteria are scarce. To our knowledge, no
463 homoplasy studies have used this method on *E. coli.*
464
465 The increase of 3GC resistance imposes a clinical threat by restricting treatment options
466 and it is essential to understand the underlying resistance mechanisms. To be able to explore
467 these mechanisms we selected primarily clinical *E. coli* strains. The current study is directed on
468 exploring AmpC mediated CTX resistance. Therefore, we included isolates that are already
469 suspected for increased AmpC production based on elevated FOX resistance. Since a random
470 sample of *E. coli* would limit finding homoplasy-based associated promoter mutations with CTX
471 resistance. A downside of these selection criteria might be that we over-estimated certain
472 genetic variants associated with the trait, as we do not know the frequency of these variants in
473 the general population. Despite the fact that the spontaneous mutation rate in *E. coli* is relatively
474 low [56], it is still likely that this particular mutation occurs often in the general population, given
475 the vast amounts of *E. coli* in the environment [57], providing ample opportunities for adaptation
476 to antibiotics and arguing for antibiotics of which genomic adaptation requires multiple
477 mutations in order to develop resistance.
478
479 Findings of this study have a number of implications for future practice. This study not
480 only grants insights into how chromosomal-encoded antibiotic resistance evolves, but also
481 provides potential strategies for future homoplasy-based association studies. Furthermore, the

482 use of genome-wide homoplasy-based analysis could be applied to optimize outbreak analysis.
483 Prior studies have optimized outbreak analysis by removing recombinant regions [58,59].
484 Homoplasy events disturbs the true phylogeny, hence, removing genomic positions which are
485 heavily affected by homoplasy could improve tree topology, thereby refining outbreak analysis,
486 although this strategy is still under debate [60].
487

488 **Conclusions**

489 To conclude, our method demonstrates extreme levels of homoplasy in *E. coli* that are
490 significantly associated with CTX resistance. Greater access to WGS data provides new
491 opportunities to perform large-scale genome-wide analysis. Homoplasy-based methods can have
492 a potential role in future studies as they constitute an effective strategy to relate phenotypic
493 traits to variable genomic positions.
494

495 **Funding information**

496 Not applicable
497

498 **Author contributions**

499 HW, JK, and MH conceived and supervised the study. JC, ED, EK, JS, JV, WM, and KN performed
500 the data acquisition. JC and ED performed the data analysis. JC performed bioinformatic
501 analysis. JC, ED, and MH performed the data interpretation and wrote the manuscript. All
502 authors read and approved the final manuscript. All authors read and approved the final
503 manuscript.

Acknowledgements

Special thanks to A. C. J. Soer (Department of Medical Microbiology and Radboudumc Center for Infectious Diseases, Radboudumc, Nijmegen, the Netherlands), B. A. Lamberts (Department of Medical Microbiology and Immunology, Rijnstate, Arnhem, the Netherlands) and C. Verhulst (Department of Infection Control, Amphia Ziekenhuis, Breda, The Netherlands and Laboratory for Microbiology, Microvida, Location Breda, The Netherlands) for handling the samples on the lab and creating the Illumina sequence libraries.

M. P. Kwint and R. Derks (Department of Human Genetics, Radboudumc, Nijmegen, the Netherlands) for helping with SMRT sequencing on the PacBio Sequel I.

Many thanks to M. Janssens (Laboratory for Medical Microbiology and Immunology, Elisabeth-Tweesteden Hospital, Tilburg, the Netherlands), S. Van Leest (Laboratory for Microbiology, Microvida, location Bravis Hospital, the Netherlands), K. T. Veldman and D. J. Mevius (department of Bacteriology and Epidemiology, Wageningen Bioveterinary Research, Lelystad, the Netherlands, E.A. Reuland (Medical Microbiology and Infection Control, Amsterdam UMC location VUmc, Amsterdam, the Netherlands), W. H. F. Goessens (Erasmus University Medical Center, Rotterdam, Netherlands), R. W. Bosboom (Department of Medical Microbiology and Immunology, Rijnstate, Arnhem, the Netherlands) and P. Vos (Check-Points, Wageningen, the Netherlands) for providing samples of which some are included in this study.

526 **Conflicts of interest**

527 The authors declare that there are no conflicts of interest.

528

529 **References**

530 1. Weinstein RA, Gaynes R, Edwards JR. Overview of Nosocomial Infections Caused by Gram-
531 Negative Bacilli. Clin Infect Dis. 2005;41:848–54.

532 2. Pitout JDD. Extraintestinal pathogenic Escherichia coli: A combination of virulence with
533 antibiotic resistance. Front. Microbiol. 2012.

534 3. Jacoby GA. AmpC B-Lactamases. Clin Microbiol Rev [Internet]. 2009;22:161–82. Available
535 from: http://cmr.asm.org/cgi/doi/10.1128/CMR.00036-08

536 4. Martinez- L, Simonsen GS. EUCAST_detection_of_resistance_mechanisms_170711. 2017;1–
537 43. Available from:
538 http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Resistance_mechanisms/EUC
539 AST_detection_of_resistance_mechanisms_170711.pdf.

540 5. Tracz DM, Boyd DA, Bryden L, Hizon R, Giercke S, Caeseele P V., et al. Increase in ampC
541 promoter strength due to mutations and deletion of the attenuator in a clinical isolate of
542 cefoxitin-resistant Escherichia coli as determined by RT-PCR. J Antimicrob Chemother.
543 2005;55:768–72.

544 6. Tracz DM, Boyd DA, Hizon R, Bryce E, McGeer A, Ofner-Agostini M, et al. ampC gene
545 expression in promoter mutants of cefoxitin-resistant Escherichia coli clinical isolates. FEMS
546 Microbiol Lett [Internet]. 2007;270:265–71. Available from:
547 https://academic.oup.com/femsle/article-lookup/doi/10.1111/j.1574-6968.2007.00672.x

548 7. De Smet AMGA, Kluytmans JAJW, Cooper BS, Mascini EM, Benus RFJ, Van Der Werf TS, et al.
549 Decontamination of the digestive tract and oropharynx in ICU patients. N Engl J Med [Internet].
550 Massachussetts Medical Society; 2009 [cited 2020 May 14];360:20–31. Available from:
551 http://www.nejm.org/doi/abs/10.1056/NEJMoa0800394
552 8. Aardema H, Bult W, Van Hateren K, Dieperink W, Touw DJ, Alffenaar JWC, et al. Continuous
553 versus intermittent infusion of cefotaxime in critically ill patients: A randomized controlled trial
554 comparing plasma concentrations. J Antimicrob Chemother [Internet]. 2020 [cited 2020 May
555 14];75:441–8. Available from: https://academic.oup.com/jac/article/75/2/441/5614359
556 9. Wake DB. Homoplasy: the result of natural selection, or evidence of design limitations? Am
557 Nat. 1991;138:543–67.
558 10. Crispell J, Balaz D, Gordon SV. Homoplasyfinder: A simple tool to identify homoplasies on a
559 phylogeny. Microb Genomics [Internet]. 2019 [cited 2019 Aug 7];5. Available from:
560 http://www.danielwilson.me.uk/
561 11. Kluge AG, Farris JS. Quantitative phyletics and the evolution of anurans. Syst Biol.
562 1969;18:1–32.
563 12. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis
564 identifies targets of convergent positive selection in drug-resistant Mycobacterium
565 tuberculosis. Nat Genet. 2013;45:1183–9.
566 13. Mortimer TD, Weber AM, Pepperell CS. Signatures of Selection at Drug Resistance Loci in
567 Mycobacterium tuberculosis. Gilbert JA, editor. mSystems [Internet]. 2018;3:1–9. Available
568 from: https://msystems.asm.org/lookup/doi/10.1128/mSystems.00108-17
569 14. Ruesen C, Chaidir L, van Laarhoven A, Dian S, Ganiem AR, Nebenzahl-Guimaraes H, et al.

570 Large-scale genomic analysis shows association between homoplastic genetic variation in
571 Mycobacterium tuberculosis genes and meningeal or pulmonary tuberculosis. BMC Genomics.
572 BioMed Central Ltd.; 2018;19.
573 15. Miotto P, Cabibbe AM, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, et al.
574 Mycobacterium tuberculosis pyrazinamide resistance determinants: A multicenter study. Nacy
575 CA, editor. MBio [Internet]. 2014;5:1–10. Available from:
576 https://mbio.asm.org/lookup/doi/10.1128/mBio.01819-14
577 16. Coolen JPM, Den Drijver EPM, Kluytmans JAJW, Verweij JJ, Lamberts BA, Soer JACJ, et al.
578 Development of an algorithm to discriminate between plasmid- and chromosomal-mediated
579 AmpC β-lactamase production in Escherichia coli by elaborate phenotypic and genotypic
580 characterization. J Antimicrob Chemother [Internet]. 2019;74:3481–8. Available from:
581 https://academic.oup.com/jac/advance-article/doi/10.1093/jac/dkz362/5554444
582 17. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new
583 genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol
584 [Internet]. 2012;19:455–77. Available from:
585 http://www.liebertpub.com/doi/10.1089/cmb.2012.0021
586 18. Beghain J, Bridier-Nahmias A, Nagard H Le, Denamur E, Clermont O. ClermonTyping: An
587 easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microb
588 Genomics. 2018;4:1–8.
589 19. Seemann T. mlst. Github https://github.com/tseemann/mlst.
590 20. Jolley KA, Maiden MCJ. BIGSdb: Scalable analysis of bacterial genome variation at the
591 population level. BMC Bioinformatics. 2010;11.

592 21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol
593 Biol [Internet]. 1990;215:403–10. Available from:
594 https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602
595 22. Peter-Getzlaff S, Polsfuss S, Poledica M, Hombach M, Giger J, Böttger EC, et al. Detection of
596 AmpC beta-lactamase in Escherichia coli: Comparison of three phenotypic confirmation assays
597 and genetic analysis. J Clin Microbiol [Internet]. 2011;49:2924–32. Available from:
598 http://jcm.asm.org/cgi/doi/10.1128/JCM.00091-11
599 23. Seemann T. Abricate [Internet]. Github; Available from:
600 https://github.com/tseemann/abricate
601 24. Caroff N, Espaze E, Gautreau D, Richet H, Reynaud A. Analysis of the effects of -42 and -32
602 ampC promoter mutations in clinical isolates of Escherichia coli hyperproducing AmpC. J
603 Antimicrob Chemother. 2000;45:783–8.
604 25. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies
605 from short and long sequencing reads. Phillippy AM, editor. PLoS Comput Biol [Internet].
606 2017;13:e1005595. Available from: https://dx.plos.org/10.1371/journal.pcbi.1005595
607 26. Tatusova T, Dicuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI
608 prokaryotic genome annotation pipeline. Nucleic Acids Res [Internet]. 2016;44:6614–24.
609 Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw569
610 27. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: An
611 update on prokaryotic genome annotation and curation. Nucleic Acids Res [Internet].
612 2018;46:D851–60. Available from: http://academic.oup.com/nar/article/46/D1/D851/4588110
613 28. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for

614 large alignments. Poon AFY, editor. PLoS One [Internet]. 2010;5:e9490. Available from:
615 https://dx.plos.org/10.1371/journal.pone.0009490
616 29. European Committee on Antimicrobial Susceptibility Testing (EUCAST). The European
617 Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs
618 and zone diameters. Version 10.0, 2020 [Internet]. 2020. Available from:
619 http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_10.0_Bre
620 akpoint_Tables.pdf
621 30. Mehta CR, Patel NR. A network algorithm for performing fisher's exact test in r × c
622 contingency tables. J Am Stat Assoc [Internet]. Taylor & Francis; 1983;78:427–34. Available
623 from: https://doi.org/10.1080/01621459.1983.10477989
624 31. Benjamini Y, Hochberg Y. <Benjamini&Hochberg1995_FDR.pdf>. J R Stat Soc Ser B
625 [Internet]. 1995;57:289–300. Available from: http://www.jstor.org/stable/2346101
626 32. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating
627 and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of
628 Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.
629 33. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. The EnteroBase user's guide, with case
630 studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic
631 diversity. Genome Res. 2020;30:138–52.
632 34. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid
633 phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using
634 Gubbins. Nucleic Acids Res [Internet]. 2015;43:e15. Available from:
635 http://academic.oup.com/nar/article/43/3/e15/2410982/Rapid-phylogenetic-analysis-of-large-

636 samples-of

637 35. Letunic I, Bork P. Interactive Tree of Life (iTOL) v4: Recent updates and new developments.

638 Nucleic Acids Res [Internet]. 2019;47:W256–9. Available from:

639 https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkz239/5424068

640 36. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator.

641 Genome Res [Internet]. 2004;14:1188–90. Available from:

642 ftp://ftp.ncbi.nih.gov/genomes/Bacteria

643 37. Connors J, Krzywinski M, Schein J, Gascoyne R, Horsman D, Jones SJ, et al. Circos : An

644 information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

645 38. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: An

646 interactive viewer for bacterial population genomics. Kelso J, editor. Bioinformatics [Internet].

647 2018;34:292–3. Available from:

648 https://academic.oup.com/bioinformatics/article/34/2/292/4212949

649 39. Crispell J, Benton CH, Balaz D, De Maio N, Ahkmetova A, Allen A, et al. Combining genomics

650 and epidemiology to analyse bi-directional transmission of mycobacterium bovis in a multi-host

651 system. Elife. 2019;8:1–36.

652 40. Van Dorp L, Gelabert P, Rieux A, De Manuel M, De-Dios T, Gopalakrishnan S, et al.

653 Plasmodium vivax Malaria Viewed through the Lens of an Eradicated European Strain. Mol Biol

654 Evol [Internet]. 2020;37:773–85. Available from:

655 https://www.biorxiv.org/content/10.1101/736702v1

656 41. Nelson EC, Gay Elisha B. Molecular basis of ampC hyperproduction in clinical isolates of

657 Escherichia coli [Internet]. Antimicrob. Agents Chemother. 1999. Available from:

658 http://aac.asm.org/

659 42. Mulvey MR, Bryce E, Boyd DA, Ofner-Agostini M, Land AM, Simor AE, et al. Molecular
660 characterization of cefoxitin-resistant Escherichia coli from Canadian hospitals. Antimicrob
661 Agents Chemother. 2005;49:358–65.

662 43. Yona AH, Alm EJ, Gore J. Random sequences rapidly evolve into de novo promoters. Nat
663 Commun [Internet]. 2018;9:1530. Available from: http://www.nature.com/articles/s41467-018-
664 04026-w

665 44. Liu S, Libchaber A. Some aspects of E. coli promoter evolution observed in a molecular
666 evolution experiment. J Mol Evol. 2006;62:536–50.

667 45. Korotkov K V., Sandkvist M, Hol WGJ. The type II secretion system: Biogenesis, molecular
668 architecture and mechanism. Nat. Rev. Microbiol. 2012. p. 336–51.

669 46. Ho TD, Davis BM, Ritchie JM, Waldor MK. Type 2 secretion promotes enterohemorrhagic
670 Escherichia coli adherence and intestinal colonization. Infect Immun. 2008;76:1858–65.

671 47. Baldi DL, Higginson EE, Hocking DM, Praszkier J, Cavaliere R, James CE, et al. The type II
672 secretion system and its ubiquitous lipoprotein substrate, SslE, are required for biofilm
673 formation and virulence of enteropathogenic Escherichia coli. Infect Immun. 2012;80:2042–52.

674 48. Kulkarni R, Dhakal BK, Slechta ES, Kurtz Z, Mulvey MA, Thanassi DG. Roles of putative type II
675 secretion and type IV pilus systems in the virulence of uropathogenic Escherichia coli. PLoS One.
676 2009;4.

677 49. Dunne KA, Chaudhuri RR, Rossiter AE, Beriotto I, Browning DF, Squire D, et al. Sequencing a
678 piece of history: Complete genome sequence of the original Escherichia coli strain. Microb
679 Genomics. 2017;3.

680 50. Vanderlinde EM, Strozen TG, Hernández SB, Cava F, Howard SP. Alterations in peptidoglycan
681 cross-linking suppress the secretin assembly defect caused by mutation of GspA in the type II
682 secretion system. J Bacteriol [Internet]. 2017 [cited 2020 Mar 25];199. Available from:
683 http://jb.asm.org/
684 51. Juan C, Torrens G, Barceló IM, Oliver A. Interplay between Peptidoglycan Biology and
685 Virulence in Gram-Negative Pathogens. Microbiol Mol Biol Rev [Internet]. 2018 [cited 2020 Mar
686 25];82. Available from: http://mmbr.asm.org/
687 52. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using
688 genome-wide association studies: A new direction for bacteriology. Genome Med. BioMed
689 Central Ltd.; 2014.
690 53. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr.
691 Opin. Microbiol. Elsevier Ltd; 2015. p. 17–24.
692 54. Shapiro BJ, David LA, Friedman J, Alm EJ. Looking for Darwin's footprints in the microbial
693 world. Trends Microbiol. 2009;17:196–204.
694 55. Benjak A, Avanzi C, Singh P, Loiseau C, Girma S, Busso P, et al. Phylogenomics and
695 antimicrobial resistance of the leprosy bacillus Mycobacterium leprae. Nat Commun. Nature
696 Publishing Group; 2018;9.
697 56. Lee H, Popodi E, Tang H, Foster PL. Rate and molecular spectrum of spontaneous mutations
698 in the bacterium Escherichia coli as determined by whole-genome sequencing. Proc Natl Acad
699 Sci U S A. 2012;109.
700 57. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal
701 Escherichia coli. Nat Rev Microbiol [Internet]. Nature Publishing Group; 2010;8:207–17.

702 Available from: http://dx.doi.org/10.1038/nrmicro2298

703 58. Escobar-Páramo P, Sabbagh A, Darlu P, Pradillon O, Vaury C, Denamur E, et al. Decreasing

704 the effects of horizontal gene transfer on bacterial phylogeny: The Escherichia coli case study.

705 Mol Phylogenet Evol [Internet]. Academic Press Inc.; 2004;30:243–50. Available from:

706 https://linkinghub.elsevier.com/retrieve/pii/S1055790303001817

707 59. Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, et al. The epidemic of

708 extended-spectrum-β-lactamase-producing Escherichia coli ST131 is driven by a single highly

709 pathogenic subclone, H30-Rx. Parkhill J, editor. MBio [Internet]. 2013;4:1–10. Available from:

710 https://mbio.asm.org/lookup/doi/10.1128/mBio.00377-13

711 60. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to

712 recombination but demographic inference is not. Vidaver AK, editor. MBio [Internet]. American

713 Society for Microbiology; 2014;5. Available from:

714 https://mbio.asm.org/lookup/doi/10.1128/mBio.02158-14

715

716 **LEGENDS**

717 **FIG 1** Schematic of workflow used to perform the homoplasy-based association analysis.

718 Starting from the top A) the *de novo* assembly of the NextSeq/MiSeq reads and B) the hybrid

719 assembly of the reference chromosome ampC_069. On the left side C) the alignment of

720 promoter/attenuator region. In the middle D) the coreSNP analysis for the phylogeny used in E)

721 the homoplasy analysis combined with F) the fullSNP data on the right, which was also used for

722 G) the statistics (Fisher Exact & FDR) to relate cefotaxime (CTX) resistance to SNP positions. H)

723 Inferring recombination events using Gubbins.

724

725 **FIG 2** Approximately maximum-likelihood phylogenetic tree of all 172 *E. coli* isolates based on
726 the coreSNP alignment with the resistance for cefotaxime (CTX), *pampC* gene presence, MLSTs,
727 phylogroups, and the alignments of the promoter and the attenuator region. Positions with a CAT
728 likelihood score ≤60% are indicated as red dots.

729

730 **FIG 3** Sequence logo with probability score for the promoter and the attenuator region. The
731 Consistency Index and the minimum number of changes on the tree per position are represented
732 below the sequence logos.

733

734 **FIG 4** Circos plot for the full chromosome of ampC_0069 (accession no. CP046396) with per
735 position the various metrics used. A) The blue colored ring represents the Consistency Index
736 results per genomic position. The two red dots indicate the -42 and -32 position on the promoter.
737 The black circle line indicates the 0.07142857 Consistency Index value. B) The ring with a red
738 background shows all positions that were significantly associated to cefotaxime (CTX) resistance
739 in all non-*pampC* harboring *E. coli* isolates. Larger bars pointing outwards indicate multiple
740 significant associated positions in a small genomic region. C) The ring with the green background
741 shows all 24 positions that have a low Consistency Index of ≤ 0.05882353 and are significantly
742 associated with CTX resistance in all non-*pampC* harboring *E. coli* isolates.

743

744 **TABLE 1** Table of the distribution of AmpC promoter and attenuator variants as well as the
745 amount of different MLST and phylogroups per grouped genotype (*pampC*, hyperproducers and
746 low-level AmpC producers).
747
748 **TABLE 2** The *n*=24 positions with a significant association with cefotaxime resistance
749 (FDR ≤0.05) and with a consistency index ≤ 0.05882353.
750
751 **FIG S1** Violin plots of the log10 Consistency Indexes of the promoter and attenuator.
752
753 **FIG S2** Distribution of the log10 Consistency Indexes of all genomic position based on the *E. coli*
754 ampC_0069 reference chromosome, compared to the log10 Consistency Indexes of the
755 promoter and attenuator region.
756
757 **FIG S3** Recombination events inferred from all 172 *E. coli* isolates by Gubbins displayed along the
758 approximately maximum-likelihood phylogenetic tree based on the coreSNP alignment.
759 Phylogroups are depicted as in FIG 2. Gubbins blocks are colored red if they are ancestral, and
760 blue if they only affect one isolate. The line graph represents the recombination prevalence along
761 the sequence. The 24 positions with a significant association with cefotaxime resistance (FDR
762 ≤0.05) and a consistency index ≤ 0.05882353 are indicated on the top of the figure. The two
763 missense mutations and *ampC* promoter region are displayed in blue.
764

765 **Table S1** Classification of n=172 *E. coli* isolates in the three genotypes (*pampC*,
766 hyperproducer, low-level AmpC producers), with the results of the MLST and phylogroups
767 stratification and the different mutations in the promoter and attenuator per isolate. Isolates
768 with a CTX MIC >2 mg/L without a confirmed *pampC* gene are depicted in **bold**.
769
770 **Table S2** SNP analysis for n=172 *E. coli* isolates according to snippy statistics.
771
772

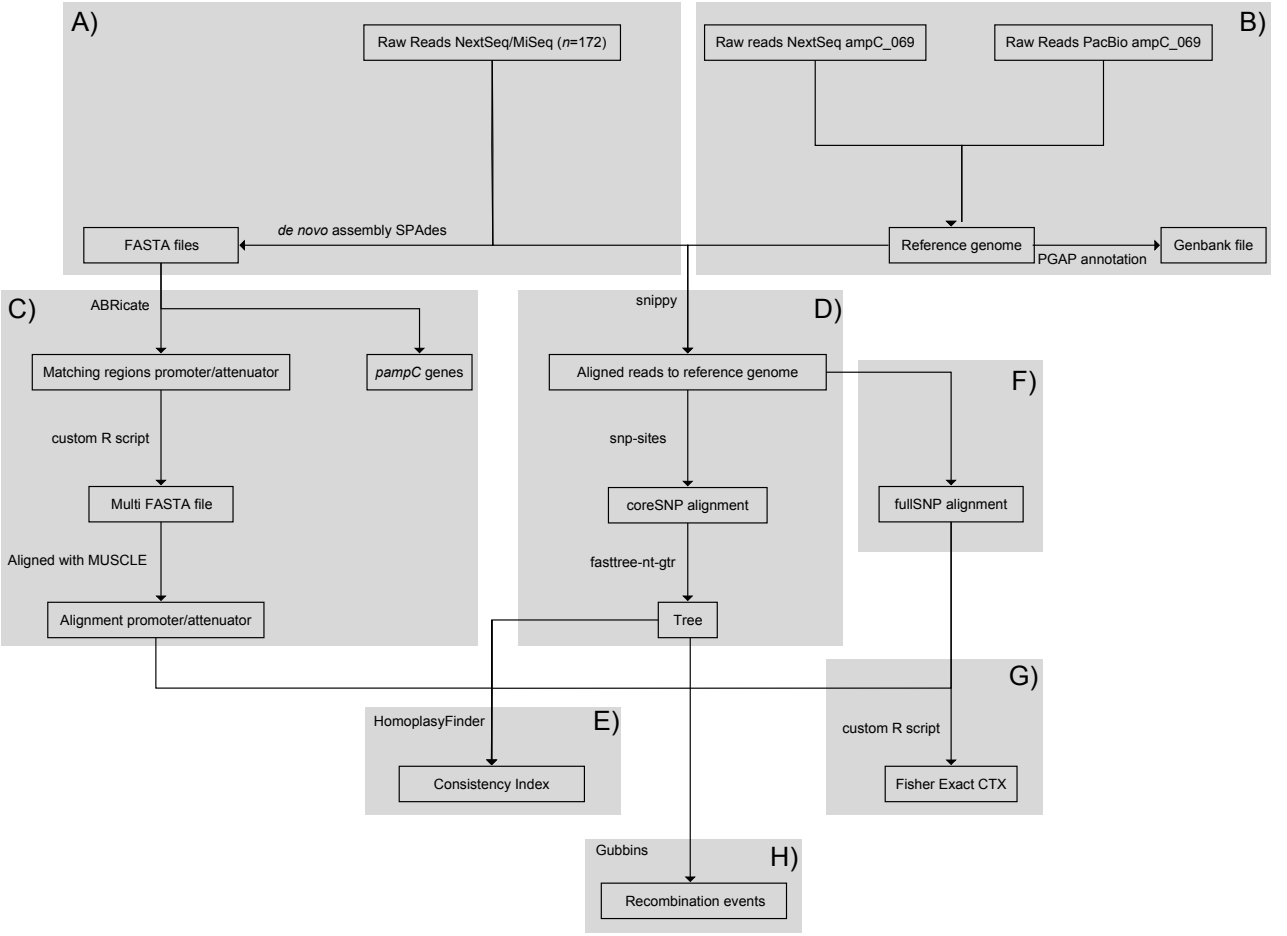

FIG 1 Schematic of workflow used to perform the homoplasy-based association analysis. Starting from the top A) the *de novo* assembly of the NextSeq/MiSeq reads and B) the hybrid assembly of the reference chromosome ampC_069. On the left side C) the alignment of promoter/attenuator region. In the middle D) the coreSNP analysis for the phylogeny used in E) the homoplasy analysis combined with F) the fullSNP data on the right, which was also used for G) the statistics (Fisher Exact & FDR) to relate cefotaxime (CTX) resistance to SNP positions. H) Inferring recombination events using Gubbins.

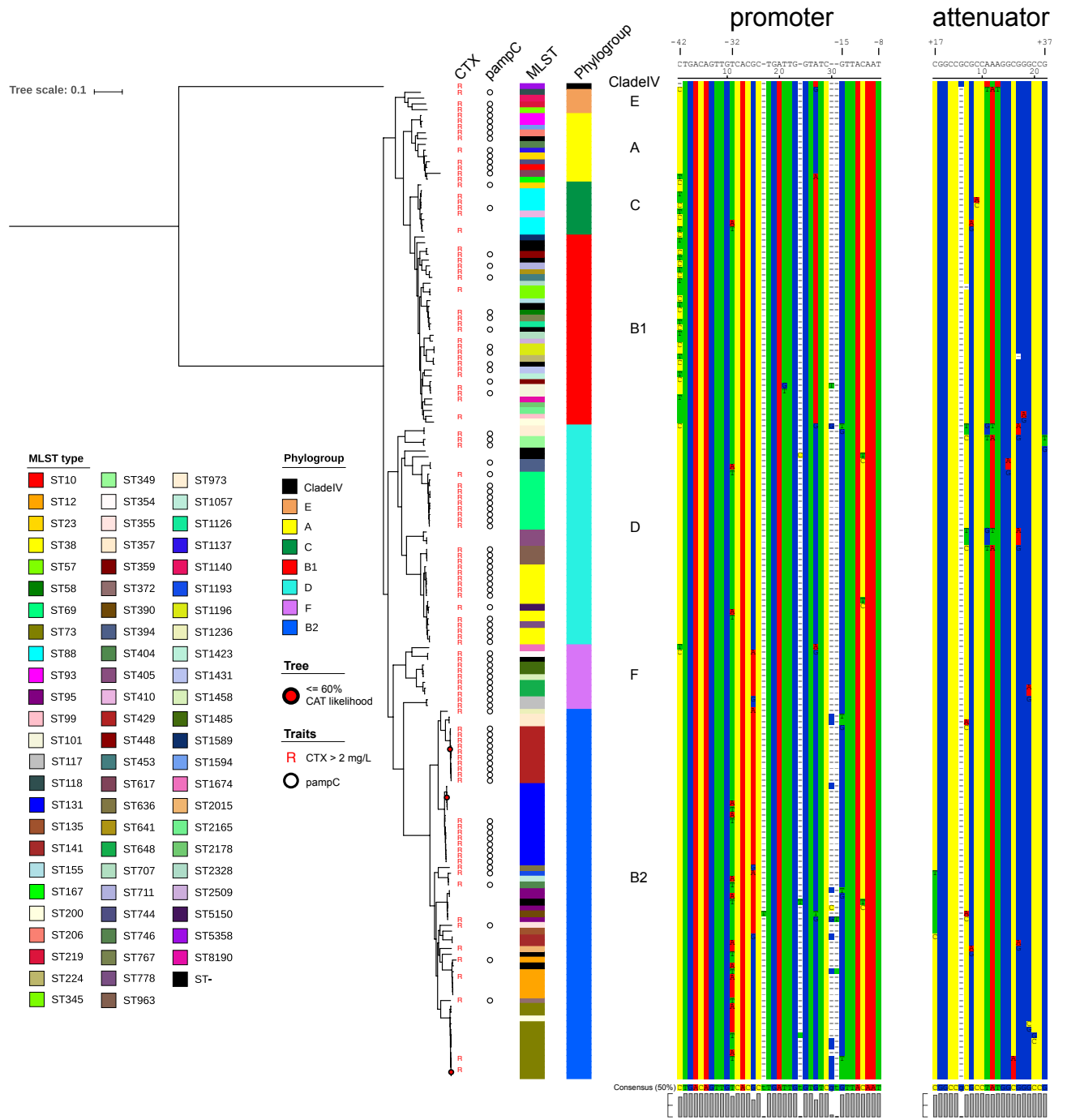

FIG 2 Approximately maximum-likelihood phylogenetic tree of all 172 *E. coli* isolates based on the coreSNP alignment with the resistance for cefotaxime (CTX), *pampC* gene presence, MLSTs, phylogroups, and the alignments of the promoter and the attenuator region. Positions with a CAT likelihood score ≤60% are indicated as red dots.

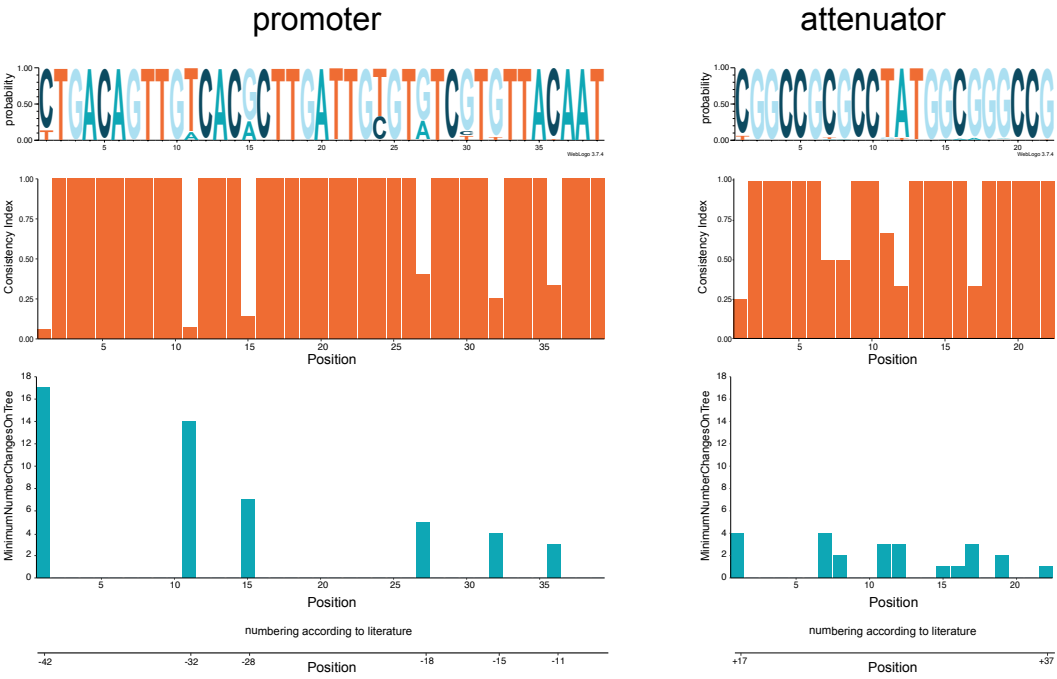

FIG 3 Sequencelogo with probability score for the promoter and the attenuator region. The Consistency Index and the minimum number of changes on the tree per position are represented below the sequence logos.

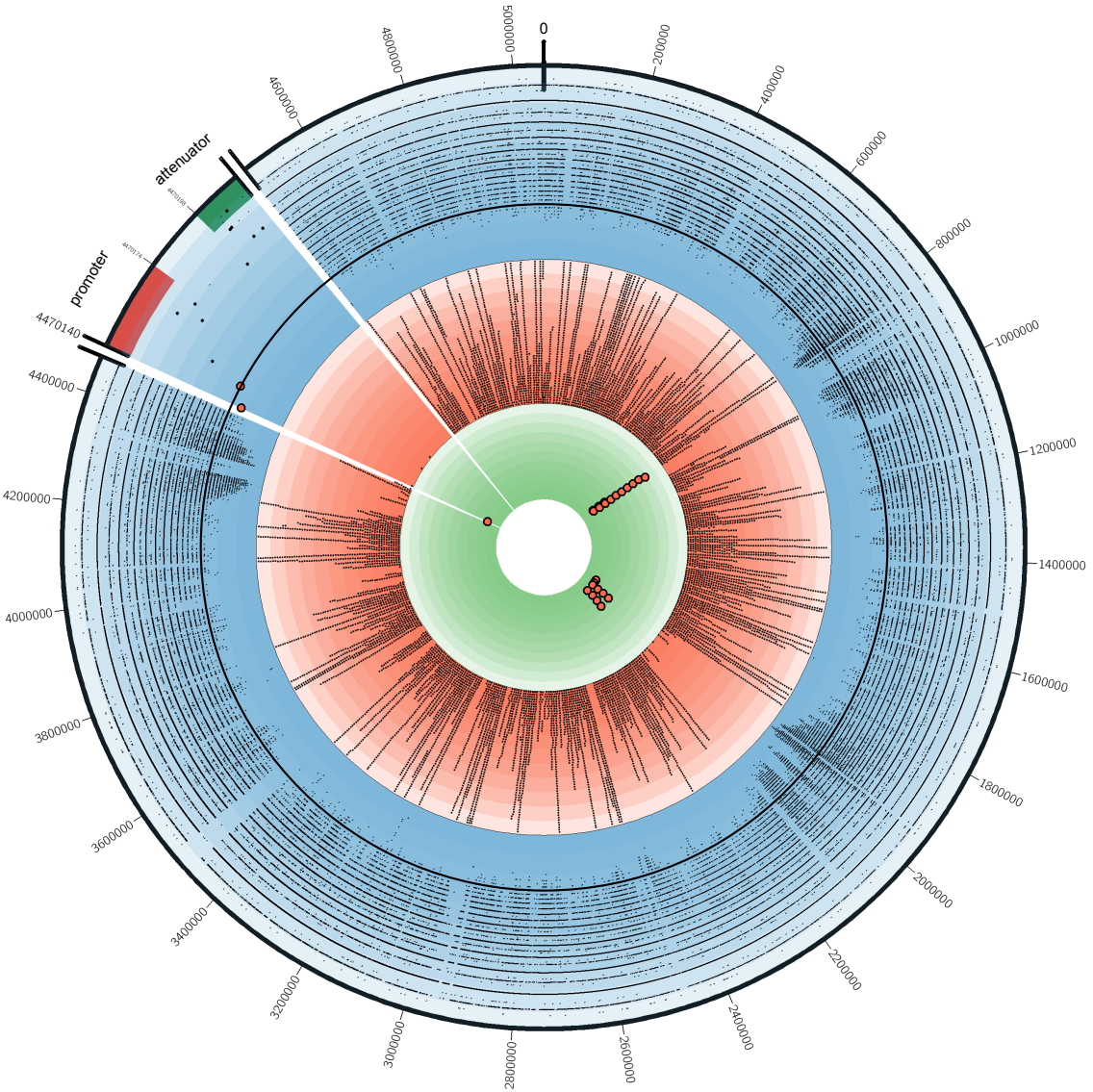

FIG 4 Circos plot for the full chromosome of ampC_0069 (accession no. CP046396) with per position the various metrics used. A) The blue colored ring represents the Consistency Index results per genomic position. The two red dots indicate the -42 and -32 position on the promoter. The black circle line indicates the 0.07142857 Consistency Index value. B) The ring with a red background shows all positions that were significantly associated to cefotaxime (CTX) resistance in all non-*pampC* harboring *E. coli* isolates. Larger bars pointing outwards indicate multiple significant associated positions in a small genomic region. C) The ring with the green background shows all 24 positions that have a low Consistency Index of ≤ 0.05882353 and are significantly associated with CTX resistance in all non-*pampC* harboring *E. coli* isolates.

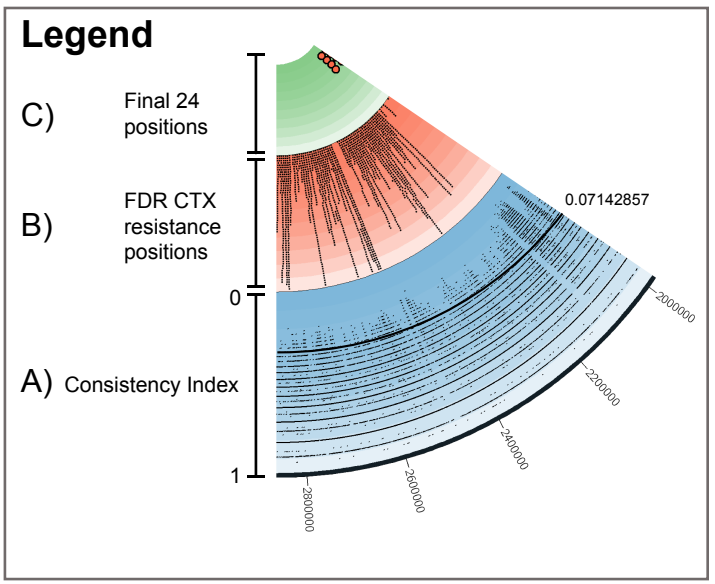

TABLE 1

Table of the distribution of AmpC promoter and attenuator variants as well as the amount of different MLST and phylogroups per grouped genotype (*pampC,* hyperproducers and low-level AmpC producers)

	Isolates	**Promoter variants**	**Attenuator variants**	**MLST**	**Phylogroups**
pampC	*n*=90	*n*=3	*n*=6	44 STs & 4 unknown	A (11.1%), B1 (13.3%), B2 (27.8%), C (2.2%), D (31.1%), E (3.3%), F (11.1%)
Hyperproducers	*n*=59	*n*=12	*n*=14	30 STs & 5 unknown	A (1.7%), B1 (30.5%), B2 (50.9%), C (8.5%), D (5.1%), F (1.7%), clade IV (1.7%)
Low-level AmpC producers	*n*=23	*n*=10	*n*=5	14 STs & 4 unknown	A (4.4%), B1 (13.0%), B2 (39.1%), C (8.7%), D (30.4%), E (4.4%)
Total	*n*=172	*n*=22	*n*=18	75 STs and 13 unknown	A (7.0%), B1 (19.2%), B2 (37.2%), C (5.2%), D (22.1%), E (2.3%), F (6.4%), clade IV (0.6%)

TABLE 2 The *n*=24 positions with a significant association with cefotaxime resistance (FDR ≤0.05) and with a consistency index ≤ 0.05882353.

Genomic position	Consistency Index	A FDR	C FDR	G FDR	T FDR	Counts ACGT	# strains	Min No Changes on Tree	Gene name	Enterobase core / accessory gene	Variant	HGVS annotation	Genomic start position	Genomic stop position	Product	Locus_tag
810581	0.05555556	0.06837015	1.00000000	0.03945241	1.00000000	41:0:122:0	163	18	*glcE*	accessory	synonymous		810312	811364	Glycolate oxidase subunit glcE	GNX12_03825
810791	0.05882353	1.00000000	1.00000000	0.04099492	0.03471297	0:0:54:109	163	17	*glcE*	accessory	synonymous		810312	811364	Glycolate oxidase subunit glcE	GNX12_03825
815680	0.05555556	1.00000000	0.14069566	1.00000000	0.04226791	0:69:0:93	162	18	*glcA*	not available	synonymous		815555	817237	Glycolate permease glcA	GNX12_03845
824522	0.05555556	1.00000000	0.03648231	1.00000000	0.14069566	0:82:0:73	155	18	*gspC*	accessory	synonymous		823719	824678	Type II secretion system protein gspC	GNX12_03865
828695	0.05555556	0.04226791	1.00000000	0.03394030	1.00000000	79:0:74:0	153	18	*gspF*	accessory	synonymous		828261	829484	Type II secretion system protein gspF	GNX12_03880
830684	0.05555556	1.00000000	0.03394030	1.00000000	0.03648231	0:80:0:73	153	18	*gspI*	accessory	synonymous		830520	830891	Type II secretion system protein gspI	GNX12_03895
830708	0.05263158	0.04099492	1.00000000	0.03648231	1.00000000	68:0:85:0	153	19	*gspI*	accessory	synonymous		830520	830891	Type II secretion system protein gspI	GNX12_03895
830732	0.05882353	1.00000000	0.03648231	1.00000000	0.04099492	0:82:0:71	153	17	*gspI*	accessory	synonymous		830520	830891	Type II secretion system protein gspI	GNX12_03895
831564	0.05882353	0.03648231	1.00000000	0.08198761	1.00000000	71:0:83:0	154	17	*gspK*	accessory	synonymous		831490	832467	General secretion pathway protein gspK	GNX12_03905
832152	0.05555556	1.00000000	0.03394030	1.00000000	0.05525251	0:86:0:72	158	18	*gspK*	accessory	synonymous		831490	832467	General secretion pathway protein gspK	GNX12_03905
832287	0.05882353	0.05525251	1.00000000	1.00000000	0.03394030	81:0:0:79	160	17	*gspK*	accessory	synonymous		831490	832467	General secretion pathway protein gspK	GNX12_03905
833451	0.05000000	0.04099492	1.00000000	1.00000000	0.03394030	64:0:0:98	162	20	*gspL*	accessory	missense	p.Ser330Thr	832464	833642	Type II secretion system protein gspL	GNX12_03910
843887	0.05263158	0.03648231	1.00000000	1.00000000	1.00000000	64:0:40:0	104	19	unnamed	accessory	synonymous		842822	844849	Capsular polysaccharide biosynthesis protein	GNX12_03950
1863004	0.05555556	0.03648231	1.00000000	0.03648231	1.00000000	126:0:46:0	172	18	*thiM*	core	missense	p.Thr122Ala	1862641	1863429	Hydroxyethylthiazole kinase	GNX12_08695
1911551	0.05882353	1.00000000	0.03648231	1.00000000	0.45554848	0:116:0:44	160	17	unnamed	accessory	synonymous		1910691	1911830	Polysaccharide export protein Wza	GNX12_08905
1946016	0.04347826	0.67473250	1.00000000	0.04099492	1.00000000	58:0:91:0	149	23	*ugd*	accessory	synonymous		1944883	1946049	UDP-glucose 6-dehydrogenase	GNX12_09060
1946067	0.04545455	1.00000000	1.00000000	1.00000000	0.04962540	0:0:65:84	149	22	non-coding region	not available	upstream		1946049	1946196	None	GNX12_09065
1946072	0.04545455	0.04962540	1.00000000	1.00000000	1.00000000	84:0:0:65	149	22	non-coding region	not available	upstream		1946049	1946196	None	GNX12_09065
1952745	0.03571429	0.05907658	1.00000000	0.03732584	1.00000000	131:0:40:0	171	28	*hisD*	accessory	synonymous		1952169	1953473	Histidinol dehydrogenase	GNX12_09100
2051214	0.05263158	0.04099492	1.00000000	1.00000000	0.09170046	137:0:0:34	171	19	unnamed	accessory	synonymous		2050846	2052204	putative sensor-like heavy metal sensor histidine kinase	GNX12_09550
2051220	0.05263158	0.04099492	1.00000000	1.00000000	0.09170046	137:0:0:34	171	19	unnamed	accessory	synonymous		2050846	2052204	heavy metal sensor histidine kinase	GNX12_09550
2057518	0.05263158	0.04099492	1.00000000	0.03648231	1.00000000	111:0:60:0	171	19	*dcm*	core	synonymous		2056604	2058022	DNA-cytosine methyltransferase	GNX12_09575
2068593	0.05882353	0.03732584	1.00000000	0.11254604	1.00000000	45:0:123:0	168	17	*fliM*	accessory	synonymous		2067810	2068814	Flagellar motor switch protein fliM	GNX12_09650
4470140	0.05882353	1.00000000	0.03394030	1.00000000	0.03394030	0:147:0:24	171	17	*ampC* promoter	core	upstream		4470140	4470174	None	GNX12_21360