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 (word limit: 250) (currently 242)

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. To study this, 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 association of the -42 C>T mutation with CTX resistance was confirmed to be significant (FDR < 0.05). Furthermore, this mutation was only observed in *E. coli* lacking a plasmid-encoded *ampC* gene. Our findings support that CTX resistance in *E. coli* is caused by recurrent mutations in the *ampC* promoter region or by obtaining a plasmid that harbors *ampC* β-lactamase genes.

IMPORTANCE (word limit: 150) (currently 144)

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

47 as well. In *E. coli* chromosome-encoded β-lactam resistance can be caused by alterations in the
48 promoter region of the *ampC* gene. To improve our understanding of how frequently these
49 alterations occur, a comprehensive interpretation of the evolution of these mutations is
50 essential. This study is the first to apply genome-wide homoplasy analysis to better perceive
51 adaptation of the *E. coli* genome to antibiotics. Thereby, this study grants insights into how
52 chromosomal-encoded antibiotic resistance evolves and, by combining genome-wide
53 association studies with homoplasy analyses, provides potential strategies for future
54 association studies into the causes of antibiotics resistance.
55
56 **INTRODUCTION**
57 *Escherichia coli* is an important pathogen in both community and healthcare-associated
58 infections (1, 2). In the past decades, a substantial increase in resistance to third-generation
59 cephalosporin (3GC) antibiotics in *E. coli* has been observed worldwide, mainly caused by the
60 production of extended-spectrum β-lactamases (ESBL) and AmpC β-lactamases, restricting
61 available treatment options for common infections (3). AmpC β-lactamases differ from ESBL as
62 they not only hydrolyze broad-spectrum penicillins and cephalosporins, but also cephamycins.
63 Moreover, AmpC β-lactamases are not inhibited by ESBL-inhibitors like clavulanic acid (3),
64 limiting antibiotic treatment options even further. A widely used method to screen for AmpC
65 production is the use of cefoxitin (FOX), a member of the cephamycins (4).
66 Although *ampC* β-lactamase genes can be plasmid-encoded (*pampC*), they are also
67 encoded on the chromosomes of numerous Enterobacterales. *E. coli* naturally carries a
68 chromosomal-mediated *ampC* (*campC*) gene, but unlike most other Enterobacterales this gene

69 is non-inducible due to the absence of the *ampR* regulator gene (3). Consequently,
70 chromosomal AmpC production in *E. coli* is exclusively regulated by promoter and attenuator
71 mechanisms. This results in constitutive low-level c*ampC* expression that still allows the use of
72 3GC antibiotics, such as cefotaxime (CTX), to treat *E. coli* infections (3). However, various
73 mutations in the promoter/attenuator region of *E. coli* may cause constitutive hyperexpression
74 of *campC* (5, 6), thereby increasing the MICs for broad-spectrum penicillins and cephalosporins
75 and limiting appropriate treatment options.

76 A wide variety of promoter and attenuator mutations have been related to AmpC
77 hyperproduction (6). AmpC hyperproduction is primarily caused by alterations of the *ampC*
78 promoter region, leading to a promoter sequence that more closely resembles the *E. coli*
79 consensus sigma 70 promoter with a TTGACA -35 box separated by 17 bp from a TATAAT -10
80 box. These alterations can be divided into different variants associated with e.g. an alternate
81 displaced promoter box, a promotor box mutation or an alternate spacer length due to
82 insertions (6). Furthermore, mutations of the attenuator sequence can lead to changes in the
83 hairpin structure that strengthen the effect of promoter alterations on AmpC hyperproduction.
84 In a study on cefoxitin-resistant *E. coli* isolated from Canadian hospitals, Tracz *et al.* described
85 52 variants of the promoter and attenuator region (6). In this study a two-step quantitative
86 reverse transcriptase (qRT-)PCR was used to determine the effect of promoter/attenuator
87 variants on *ampC* expression. Various mutations were related to different delta–delta cycle
88 threshold values in the RT-PCR and corresponding variations in FOX resistance. An interesting
89 observation that emerged from this study was that the -32T>A and the -42C>T mutation were
90 the major alterations that strengthened the *ampC* promoter. Both result in a consensus -35

91 box. Although it is known that AmpC production leads to FOX resistance as studied by Tracz *et*
92 *al.* the effect of various mutations on resistance to a 3GC antibiotic such as CTX have not been
93 explored. This is relevant because CTX is commonly used in the treatment of patients with
94 severe *E. coli* infections, often in combination with selective digestive tract decontamination in
95 Intensive Care Units (7, 8).

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

110 In the present study, we hypothesize that some of the mutations in the *ampC*
111 promoter/attenuator region are homoplastic and are associated with CTX resistance. To test
112 our hypothesis, we performed genome-wide homoplasy analysis and combined it with a

113 genome-wide analysis of polymorphisms associated with CTX resistance by constructing a
114 tailored *E. coli* reference chromosome and combining it with WGS data of 172 FOX resistant *E.*
115 *coli* isolates previously collected by our research group (16).
116
117 **RESULTS**
118 **A short description of the *E. coli* collection used in this study.** To study genetic
119 homoplasy events in suspected AmpC producing *Escherichia coli,* FOX MIC > 8 ug/ml and ESBL
120 phenotype negative *E. coli* isolates (*n*=172) were selected as previously described by Coolen *et*
121 *al.* (16) (see Table S1 in supplemental material). The entire collection was subjected to whole-
122 genome sequencing followed by *de novo* assembly of the sequence reads to obtain contigs.
123 **Description of MLST and phylogroup variants within the *E. coli* collection.** To access
124 the genetic diversity of our *E. coli* collection, we identified both multi-locus sequence typing
125 (MLST) and phylogroup variants of each of the 172 *E. coli* isolates. Seventy-five different
126 sequence types (STs) were identified, of which ST 131 (8.1% *n*=14), ST 38 (7.0% *n*=12), and ST 73
127 (7.0% *n*=12) were the most prevalent. The sequence types of 13 isolates are unknown.
128 Phylogroup stratification revealed that the isolates belonged to eight different phylogroups
129 (Table 1). Phylogroup B1, B2, and D were the most prevalent. One isolate belonged to
130 *Escherichia* clade IV (st. no. ampC_0128).
131 **Detection and description of variants in the *ampC* promoter and attenuator.** We
132 examined the whole *E. coli* genome. However, we firstly focused on mutations in the *ampC*
133 promoter and attenuator region. Previously described mutations in the *ampC* promoter region
134 that according to Tracz *et al.* lead to "hyperproduction" of AmpC were detected in 59 (34.4%)

135 of the isolates (6). These isolates were therefore labelled as hyperproducers. Analysis of the
136 promotor area (-42 to -8) resulted in 21 different variants and the wild type (see Table 1). In the
137 attenuator region (+17 to +37), 18 different variants were identified (see Table 1). One isolate
138 (ampC_0128) showed an unusual promoter variant, a four-nucleotide deletion (-45_-
139 42delATCC). Moreover, an insertion (21_22insG) of unknown function was detected in the
140 attenuator (see Table S1 in supplemental material).

141 **Detection and description of Plasmid-mediated *ampC* variants.** As we aim to associate
142 chromosomal mutations with CTX resistance, differentiation of *pampC* harboring isolates from
143 non-*pampC* harboring isolates was required. Genomic analysis showed that 90 (52.3%) of the
144 isolates harbored a *pampC* gene of which $bla_{\text{CMY-2}}$ was the most prevalent (*n*=78). One isolate
145 harbored two different *pampC* genes ($bla_{\text{CMY-4}}$ and $bla_{\text{DHA-1}}$) (ampC_0119). One isolate
146 contained a $bla_{\text{CTX-M-27}}$ gene combined with a $bla_{\text{CMY-2}}$ gene (ampC_0114) but was ESBL disc test
147 negative (see Table S1 in supplemental material). In 23 (13.4%) of the isolates neither *pampC*
148 nor described mutations related to AmpC hyperproduction were detected and are noted as
149 low-level AmpC producers.

150 **Creating a tailored reference chromosome using Illumina and PacBio SMRT**
151 **sequencing.** To be able to reconstruct an accurate phylogeny we selected *E. coli* isolate
152 ampC_0069, one of the strains of the study, to use as reference chromosome for SNP calling.
153 The tailored reference chromosome was constructed through a hybrid assembly of *n*=4,423,109
154 2x 150 bp Illumina NextSeq 500 paired-end reads together with *n*=218,475 PacBio Sequel SMRT
155 subreads (median 5,640 bp). This resulted in a high-quality full circular chromosome of *E. coli*
156 isolate ampC_0069, with a size of 5,056,572 bp. This isolate belongs to ST648 and contains a

157 plasmid-encoded *bla*$_{CMY-42}$. The full circular chromosome was uploaded to GenBank accession
158 no. CP046396 and was used for further analysis. Genome annotation with the NCBI Prokaryotic
159 Genome Annotation Pipeline (PGAP) identified 4,720 Coding Sequences.

160 **SNP calling of *E. coli* isolates to reference chromosome.** To be able to reconstruct the
161 phylogeny and obtain SNP positions, we mapped reads of all isolates to the reference
162 chromosome *E. coli* ampC_0069, (accession no. CP046396) resulting in a coreSNP alignment
163 containing 314,200 variable core SNP positions. For further details per isolate see Table S2 in
164 supplemental material. To validate our SNP calling method we compared the ampC_0069
165 Illumina NextSeq 500 paired-end reads to the reference chromosome of ampC_0069, resulting
166 in 0 SNPs detected, supporting that the SNP calling data and method produce no false positives.

167 **Inferring a phylogenetic tree based on core SNP positions.** The coreSNP alignment was
168 used for further analysis. Figure 2 illustrates the approximately maximum-likelihood
169 phylogenetic tree of all 172 isolates based on the coreSNP alignment. The tree has a robust
170 topology as indicated by computerized adaptive testing (CAT) likelihood calculations, resulting
171 in only three positions with a value ≤60% (17). When focusing on the *ampC* promoter
172 mutations, they were most prevalent in phylogroups B1, B2, and C, although they were present
173 in all phylogroups except phylogroup E that lacked mutations in either the promoter or
174 attenuator region. Interestingly, two positions previously highlighted by Tracz *et al.,* -42 and -
175 32, are only mutated in the absence of a *pampC* gene, even in isolates with a similar MLST
176 (ST12, ST88, and ST131). The -42C>T mutation, which results in an alternate displaced promoter
177 box and therefore leads to increased resistance (6), is present in 24 isolates in 5 distinct
178 phylogroups and in 17 separate phylogenetic branches, indicating that this mutation is

179 homoplastic. Additionally, the -32T>A mutation in the promoter, previously also associated to
180 resistance (6), is present in 20 isolates in 3 distinct phylogroups and in 14 separate phylogenetic
181 branches. To quantify the level of homoplasy we calculated the Consistency Index.

182 **Computing the Consistency Index to detect genomic homoplastic mutations.** We
183 calculated the Consistency Index for all positions on the *E. coli* reference chromosome. A low
184 Consistency Index value for a position indicates a high degree of inconsistency with the
185 chromosomal phylogeny and can be calculated by HomoplasyFinder as described in earlier
186 studies (10, 18, 19). As can be observed in Figure 3, results clearly indicate that notwithstanding
187 multiple other low scoring Consistency Index positions in the promoter and attenuator, position
188 -42 (4,470,140) and -32 (4,470,150) are the lowest scoring, respectively 0.05882353 and
189 0.07142857 (see also Fig. S1 in the supplemental material). To access how extreme these
190 Consistency Index values are, we calculated the Consistency Index for all positions in the
191 chromosome (see Fig. S2 in the supplemental material). All Consistency Indexes < 1.0 are
192 plotted in the outer ring (ring A) of Figure 4. Results show that only 9,640 out of 5,056,572
193 positions (0.19%) had a Consistency Index ≤ 0.07142857 (see Figure 4 ring A, cutoff is indicated
194 by black circle). This clearly indicates that positions with low Consistency Indexes are rare, but
195 not unique. Although these 9,640 positions have a low Consistency Index, we do not yet know
196 their relation to CTX resistance.

197 **CTX resistance measurements.** Cefotaxime MIC measurements from Coolen *et al.* in
198 relation to the genotype of the *E. coli* isolates are shown in supplementary table 1. Eighty-four
199 of ninety (93.3%) *pampC* harboring *E. coli* were CTX resistant (MIC >2 mg/L) based on EUCAST
200 clinical breakpoints. Twenty-one of fifty-nine (35.6%) isolates categorized as hyperproducers

201 based on Trazc *et al.* were CTX resistant, primarily isolates with the -42 (*n*=15) or -32 mutation
202 (*n*=2). The *pampC* genes never occurred simultaneously with the -42 or -32 mutations in any of
203 these isolates. One of twenty-three (4.3%) isolates categorized as a low-level AmpC producer
204 (no *pampC* gene and no known ampC promoter mutation) tested CTX resistant and contained
205 an insertion in the spacer region (-16_-15insT), which has not been described by Tracz *et al* (6).
206 As depicted in Figure 2 the non-*pampC* strains with a phenotype of CTX > 2 mg/L were present
207 in all phylogroups, although CTX resistant isolates with the -42 or -32 mutation were
208 predominantly present in phylogroups B1, B2, and C.

209 **Association of SNPs with CTX resistance.** To be able to link *E. coli* chromosomal
210 mutations to CTX resistance we excluded all *E. coli* isolates with a plasmid containing an *ampC*
211 β-lactamase gene. The association of SNPs to CTX resistance phenotype (MIC > 2 mg/L) was
212 tested in the remaining 82 isolates using Fisher's Exact Test. After FDR correction to 0.05,
213 45,998 significant positions were found (see Figure 4 ring B). Mutation C>T on position -42 of
214 the *ampC* promoter was found to be significantly associated to CTX resistance (FDR = 0.034).
215 However, position -32 A>T was not significantly associated to CTX resistance (FDR = 1).

216 **Homoplasy-based association analysis.** Combining the outcome of the homoplasy
217 analysis with the significant CTX resistance associated positions results in genomic positions
218 associated to CTX resistance that have evolved multiple times independently. After selecting
219 the lowest scoring Consistency Index positions, ≤ 0.05882353, 24 relevant genomic positions
220 were identified that had both a low Consistency Index and a significant association with CTX
221 resistance. Most notably, one of these 24 positions is position -42. Only two mutations of those
222 24 that were located in genes were non-synonymous: a (conservative) missense mutation in the

223 type II secretion system protein L (*gspL*) gene leading to Ser330Thr alteration and a mutation in
224 the hydroxyethylthiazole kinase (*thiM*) gene resulting in a Thr122Ala alteration according to the
225 annotation of *E. coli* strain ampC_0069 (accession no. CP046396). In addition to the non-
226 synonymous mutation found on the *gspL* gene, eight synonymous mutations are also located in
227 genes annotated as being part of the type II secretion system. A complete overview is
228 presented in Table 2.

229 **Recombination analysis.** To verify if the level of homoplasy could be a result of
230 recombination, we used Genealogies Unbiased By recomBinations In Nucleotide Sequences
231 (Gubbins) algorithm to predict recombination events in our isolate collection (20). This analysis
232 showed frequent recombination events in our 172 *E. coli* isolates (see Fig. S3 in the
233 supplemental material). Results illustrate that recombination blocks cover the region of the
234 *gspL* and the *thiM* gene and their high homoplasy levels could thus be due to recurrent
235 recombination rather than independent mutations. Nonetheless, position -42 in the *ampC*
236 promoter is not located in a region effected by recombination as shown in Fig. S3. Moreover,
237 when inferring the phylogenetic tree corrected for recombination events as obtained from the
238 Gubbins analysis, the -42C>T mutation actually occurred, in 18 independent branches rather
239 than the 17 branches in the uncorrected tree. This supports our previous results that this
240 mutation is homoplastic, and not the results of a recurrent recombination event.

241

242 **DISCUSSION**

243 We present a genome-wide analysis in which homoplastic mutations are associated with
244 antibiotic resistance in *E. coli*. By comparing whole-genome sequencing data of 172 *E. coli*

245 isolates to a tailored reference chromosome we were able to reconstruct the evolution of the
246 genomes and therewith map recurrent events, allowing us to detect homoplasy associated to
247 CTX resistance.

248 Our foremost finding is the significant association of the -42C>T mutation, in the *ampC*
249 promoter, to CTX resistance that evolved independently at least 17 times in 5 distinct
250 phylogroups. The -42C>T mutation has been confirmed in former studies to result in AmpC
251 hyperproduction in *E. coli*. Nelson *et al.* demonstrated an 8 to 18 times increase in activity of
252 AmpC when cloning the promoter upstream a *lac* operon (21). Vice versa, Caroff *et al.* found a
253 decrease in expression of AmpC when cloning the promoter with a -42T>C mutation in a
254 pKK232-8 reporter plasmid with chloramphenicol acetyltransferase (CAT) gene (22). Tracz *et al*.
255 confirmed that the -42C>T mutation has the strongest effect on the *ampC* promoter, resulting
256 in a high expression of the *ampC* gene as detected by RT-qPCR (6). Despite the fact that the -
257 42C>T mutation has such a strong effect on AmpC production the effect of the mutation on CTX
258 MICs had not been confirmed. Moreover, the contribution of convergent evolution on this
259 position relative to the role of the expansion of a clone with a beneficial mutation at this
260 position has not been determined. That being the case, this study provides evidence that this -
261 42 C>T mutation is not a result of a recombination event and most likely evolved many times
262 independently. Remarkably, we observed that the -42C>T mutation never occurs in the
263 presence of a *pampC* gene (in zero out of twenty-four cases). This was even noticed in isolates
264 with the same MLST, i.e. ST88 -42C>T (*n*=3) and *pampC* (*n*=1), suggesting preferred exclusivity
265 for one of the resistance mechanisms. One study mentioned the co-occurrence of the -42C>T
266 mutation and a *pampC* gene in only one out of thirty-six strains (23). One could speculate that

267 the exclusivity is a matter of what arrives first, the plasmid or the mutation, after which there is
268 no selective advantage for the second mechanism, or that there is actually a fitness cost to
269 having both the mutation and the plasmid relative to having only the mutation or the plasmid.
270
271 The study performed by Tracz *et al.* showed that position -32T>A on the promotor of
272 *ampC* associates with AmpC hyperproduction that results in elevated MIC levels for FOX (6).
273 Surprisingly, in the current study no significant association of -32T>A with CTX resistance was
274 noticed despite its low Consistency Index. Only two out of twenty isolates with the -32T>A were
275 CTX resistant, four out of twenty showed an intermediate elevated CTX MIC, and fourteen were
276 susceptible for CTX. Although we do not know under which conditions this mutation did arise, it
277 can be speculated that the high level of homoplasy at the -32 position is associated with a
278 different trait, e.g. resistance against another antibiotic. Moreover, other resistance
279 mechanisms, as for instance differences in porin expression (24) or variants in AmpC protein
280 composition (25) could have an effect on 3GC susceptibility.
281
282 Prior studies discovered the importance of mutations in the promoter elements.
283 Random sequences can even evolve expression comparable to the wild-type promoter
284 elements after only a single mutation (26). Furthermore, these promotor elements evolve to
285 only a few forms indicating convergent evolution (27), as also observed in the present study. All
286 encountered variants seem to result in a sequence that resembles the *E. coli* consensus sigma
287 70 promoter more than the wild type sequence they are derived from (6).
288

289 Next to the -42C>T promoter mutation we detected twenty-three other positions in our
290 analysis that are associated with CTX resistance and have extreme high levels of homoplasy.
291 Most of these are synonymous mutations, with only two missense mutations (*thiM* and *gspL*)
292 found. It is remarkable that one missense mutation (p.Ser330Thr) is located in *gspL* that
293 encodes for a protein of the type II secretion system. The type II secretion system is used by
294 many gram-negative bacteria to translocate folded proteins from the periplasm, through the
295 outer membrane, into the extracellular milieu (28). The system is composed of 12–15 different
296 general secretory pathway (Gsp) proteins and is related to virulence of various pathogenic *E.*
297 *coli,* e.g. EHEC and UPEC (29–31). It could be that in our selection of mainly clinical samples a
298 certain predilection has occurred towards isolates with particular virulence traits. The *gspL*
299 gene has been described as being part of the accessory genome of *E. coli* (32). Our study
300 supports that this finding as some strains did not harbor this gene. Additionally, we found
301 evidence that recombination events in the type II secretion system could be the underlying
302 cause of the extreme homoplasy levels. Still, it is remarkable that missense mutation
303 p.Ser330Thr in the *gspL* gene correlates with the CTX resistance trait even though it is most
304 likely caused by a recombination event. To the best of our knowledge no relationship between
305 type II secretion system and CTX resistance has been observed before. One could hypothesize
306 that the mutation is a secondary adaptation needed to cope with the elevated AmpC
307 production, as the peptidoglycan (PG) layer is effected by AmpC hyperproduction and the type
308 II secretion system contains proteins that are partly localized in the periplasm (33, 34).
309

310 The use of genomic data to detect homoplasy events is not an uncommon scientific
311 technique (35–37). In *Mycobacterium tuberculosis*, it is a well-known method to identify
312 advantageous mutations, as they are likely to be associated with phenotypes such as drug
313 resistance, heightened transmissibility, or host adaptation (12–15). A similar approach was
314 taken recently by Benjak *et al.* to screen for highly polymorphic genes and genomic regions of
315 *Mycobacterium leprae* (38). Homoplasy-based association analysis limits phylogenetic bias by
316 correcting for genetic relatedness of strains with the same phenotype, thereby increasing
317 statistical power to find true associations (14). Taking this into account, the use of homoplasy-
318 based association analysis seems viable to relate polymorphic sites to phenotypic traits in
319 bacteria. Still, studies on other genera than mycobacteria are scarce. To our knowledge, no
320 homoplasy studies have used this method on *E. coli.*
321
322 The increase of 3GC resistance imposes a clinical threat by restricting treatment options
323 and it is essential to understand the underlying resistance mechanisms. To be able to explore
324 these mechanisms we selected primarily clinical *E. coli* strains. The current study is directed on
325 exploring AmpC mediated CTX resistance. Therefore, we included isolates that are already
326 suspected for increased AmpC production based on elevated FOX resistance. Since a random
327 sample of *E. coli* would limit finding homoplasy-based associated promoter mutations with CTX
328 resistance. A downside of these selection criteria might be that we over-estimated certain
329 genetic variants associated with the trait, as we do not know the frequency of these variants in
330 the general population. Despite the fact that the spontaneous mutation rate in *E. coli* is
331 relatively low (39), it is still likely that this particular mutation occurs frequently in the general

332 population, given the vast amounts of *E. coli* in the environment (40), providing ample
333 opportunities for adaptation to antibiotics and arguing for antibiotics of which genomic
334 adaptation requires multiple mutations in order to develop resistance.
335
336 Findings of this study have a number of implications for future practice. This study not
337 only grants insights into how chromosomal-encoded antibiotic resistance evolves, but also
338 provides potential strategies for future homoplasy-based association studies. Furthermore, the
339 use of genome-wide homoplasy-based analysis could be applied to optimize outbreak analysis.
340 Prior studies have optimized outbreak analysis by removing recombinant regions (41, 42).
341 Homoplasy events disturbs the true phylogeny, hence, removing genomic positions which are
342 heavily affected by homoplasy could improve tree topology, thereby refining outbreak analysis,
343 although this strategy is still under debate (43).
344
345 To conclude, our method demonstrates extreme levels of homoplasy in *E. coli* that are
346 significantly associated with CTX resistance. Greater access to WGS data provides new
347 opportunities to perform large-scale genome-wide analysis. Homoplasy-based methods can
348 have a potential role in future studies as it is an effective strategy to relate phenotypic traits to
349 variable genomic positions.
350
351 **MATERIALS AND METHODS**
352 **Isolate selection, DNA isolation, library preparation and DNA sequencing.** One
353 hundred seventy-two *Escherichia coli* isolates previously used by our study group (16) were

354 selected in the present study (see Table S1 in supplemental material). To summarize the
355 method; DNA isolation was performed as previously described (16), library preparations were
356 performed using Illumina Nextera XT library preparation kit (Illumina, San Diego, CA, USA), and
357 DNA sequencing was performed using an Illumina NextSeq 500 (Illumina, San Diego, CA, USA) to
358 generate 2x 150bp paired-end reads or 2x 300 bp reads on an Illumina MiSeq (Illumina, San
359 Diego, CA, USA). *De novo* assembly was also performed identical to the method as described in
360 Coolen *et al.* 2019 (16) using SPAdes version 3.11.1 (44).

361 **Phylogroup and MLST.** Phylogroup stratification was performed using ClermonTyping
362 version 1.4.0 (45). MLST STs were derived from the contigs using mlst version 2.5 pubMLST, 31
363 October 2017 (46, 47).

364 **Obtaining the *ampC* promoter/attenuator region.** To detect the promoter/attenuator
365 region a custom blast database (48) was created using the 271 bp fragment as described by
366 Peter-Getzlaff *et al.*(49) using *Escherichia coli* K-12 strain ER3413 (accession: CP009789.1)
367 ABRicate version 0.8.9 (50) was used to locate matching regions per sample and were extracted
368 and converted into multi-fasta format using a custom python script. Strains were labelled AmpC
369 hyperproducer when promoter mutations were found, as reported by Caroff *et al.* (22) and
370 Tracz *et al.* (6).

371 **Plasmid-mediated ampC detection.** Detection of *pampC* genes was performed by using
372 ABRicate version 0.5 (50) and ResFinder database (2018-02-16) as described by Coolen *et al.*
373 (16).

374 **PacBio single molecule real-time (SMRT) sequencing of *E. coli* isolate.** For PacBio SMRT
375 sequencing, genomic DNA (gDNA) was extracted using the Bacterial gDNA Isolation Kit (Norgen

376 Biotek Corp., CAN, ON, Thorold). A single *E. coli* isolate was subjected for DNA shearing using
377 Covaris g-TUBEs (Covaris Inc, US, MA, Woburn) for 30 seconds on 11,000 RPM (*g*). Each DNA
378 sample was separated into two aliquots. Size selection was performed using a 0.75% agarose
379 cassette and marker S1 on the BluePippin (Sage Science Inc, US, MA, Beverly) to obtain either 4-
380 8 kb or 4-12 kb DNA fragments. This size selection was chosen to maintain all DNA fragments
381 including these originating from plasmids (data not used in this study). Library preparation was
382 performed using the SMRTbell Template prep kit 1.0 (Pacific Biosciences, US, CA, Menlo Park).
383 For cost-effectiveness, samples were barcoded and pooled with other samples that are not
384 relevant for this study. Sequencing was conducted using the PacBio Sequel I (Pacific
385 Biosciences, US, CA, Menlo Park) on a Sequel SMRT Cell 1M v2 (Pacific Biosciences, US, CA,
386 Menlo Park) with a movie time of 10 h (and 186 min pre-extension time). Subreads per sample
387 were obtained by extracting the bam files using SMRT Link version 5.1.0.26412 (Pacific
388 Biosciences, US, CA, Menlo Park).

389 **Chromosomal reconstruction using *de novo* hybrid assembly.** To obtain a full-length
390 chromosome, Unicycler version 0.4.7 (51) (settings: --mode bold) was used, combining Illumina
391 NextSeq 500 2x 150 bp paired-end reads with PacBio Sequel SMRT subreads. Because unicycler
392 requires fasta reads as input, the subreads in bam format were converted to fasta by using
393 bam2fasta version 1.1.1 from pbbioconda (https://github.com/PacificBiosciences/pbbioconda)
394 prior to *de novo* hybrid assembly. The full circular chromosome was uploaded to NCBI and
395 annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) version 4.10 (52,
396 53).

397 **SNP analysis using *E. coli* reference ampC_0069.** Alignment of Illumina reads and SNP
398 calling was performed for all isolates to the reference chromosome of *E. coli* isolate ampC_0069
399 using Snippy version 4.3.6 (https://github.com/tseemann/snippy). A full-length alignment
400 (fullSNP) and a coreSNP alignment containing SNP positions shared among all isolates were
401 generated by using snippy-core version 4.3.6 (https://github.com/tseemann/snippy).

402 **Inferring of phylogeny.** A phylogenetic tree was inferred by using the coreSNP
403 alignment as input for FastTree(MP) version 2.1.3 SSE3 (settings: -nt -gtr) (17).

404 **Detection of Homoplasy.** The Consistency Index for all nucleotide positions on the
405 chromosome was calculated using HomoplasyFinder version 0.0.0.9000 (10). The coreSNP
406 phylogeny was used as true phylogeny and the Consistency Index was calculated using the
407 multiple sequence alignments fullSNP alignment as input.

408 **Relate mutations to CTX resistance.** To assess if certain mutations were linked to CTX
409 resistance all non-plasmid harboring *ampC E. coli* isolates were used. CTX resistance was
410 defined using EUCAST guidelines standards of CTX MIC > 2 mg/L (54). CTX MIC results were
411 obtained from our previous study (16). For each nucleotide position on the reference
412 chromosome the number of Resistant and Sensitive isolates were counted and tested for
413 adenine vs all other nucleotides, thymine vs all other nucleotides, cytosine vs all other
414 nucleotides, and guanine vs all other nucleotides creating a contingency table and performing a
415 Fisher Exact Test in R 3.6.1 (55). To correct for multiple testing, *P* values were adjusted using
416 FDR (56).

417 **Selection of genomic positions of interest.** By combining previous metrics most
418 relevant genomic positions were selected. Criteria for selection are, FDR <= 0.05 to CTX and

419 Consistency Index of <= 0.05882353. Annotation of mutation positions was obtained by using
420 the genome annotation of reference chromosome (GenBank accession no. CP046396) and
421 applying snpEff (version 4.3t) (57). The EnteroBase core-genome MLST and whole-genome
422 MLST schemes were used to distinguish core and accessory genes (58).

423 **Recombination analysis.** Gubbins version 2.4.1. (settings: -f 30) was used to detect
424 recombination regions with coreSNP alignment and tree as input (20).

425 **Visualization of data.** The interactive Tree of Life web-based tool iTOL version 5.3 was
426 used to visualize the phylogenetic tree (59). Information about CTX resistance, presence of the
427 *pampC* gene, *campC* hyperproduction as defined, MLST and phylogroup, as well as alignments
428 of promoter and attenuator region were incorporated into visualization. The sequence logo of
429 the promoter and the attenuator alignment were generated using the web-based application
430 Weblogo version 3.7 (60) (http://weblogo.threeplusone.com). A chromosome ideogram of the
431 *E. coli* isolate ampC_0069 reference chromosome was visualized using CIRCOS software
432 package version 0.69-8 (61). Consistency Index scores and significant mutations associated with
433 CTX resistance were plotted in the ideogram. Gubbins results were displayed by using
434 phandango (62).

435 **Overview of method.** A workflow graph of the methods is visualized in Fig 1 using the
436 web-based application yEd Live version 4.4.2 (https://www.yworks.com/yed-live/).

437 **Data availability.** Raw Illumina sequencing data of 171 *E. coli* isolates in this study are
438 available from the Sequence Read Archive database under accession no. SAMN15052485 to
439 SAMN15052655. Full reference chromosome of ampC_0069 is available via GenBank accession
440 no. CP046396.

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.

463 **REFERENCES**

464 1. Weinstein RA, Gaynes R, Edwards JR. 2005. Overview of Nosocomial Infections Caused by
465 Gram-Negative Bacilli. Clin Infect Dis 41:848–854.
466 2. Pitout JDD. 2012. Extraintestinal pathogenic Escherichia coli: A combination of virulence
467 with antibiotic resistance. Front Microbiol.
468 3. Jacoby GA. 2009. AmpC B-Lactamases. Clin Microbiol Rev 22:161–182.
469 4. Martinez- L, Simonsen GS. 2017. EUCAST_detection_of_resistance_mechanisms_170711
470 1–43.
471 5. Tracz DM, Boyd DA, Bryden L, Hizon R, Giercke S, Caeseele P V., Mulvey MR. 2005.
472 Increase in ampC promoter strength due to mutations and deletion of the attenuator in a
473 clinical isolate of cefoxitin-resistant Escherichia coli as determined by RT-PCR. J
474 Antimicrob Chemother 55:768–772.
475 6. Tracz DM, Boyd DA, Hizon R, Bryce E, McGeer A, Ofner-Agostini M, Simor AE, Paton S,
476 Mulvey MR, Conly J, Dow G, Embil J, Embree J, Frenette C, Gardam M, Gourdeau M, John
477 M, Henderson BA, Hutchinson J, Ishak M, Khan A, Kibsey P, Johnston L, Langley J, Loeb M,
478 Matlow A, Michaud S, Miller M, Moore D, Roth V, Suh K, Taylor G, Thomas E, Tremblay C,
479 Vearncombe M, Weiss K, Wong A, Zoutman D. 2007. ampC gene expression in promoter
480 mutants of cefoxitin-resistant Escherichia coli clinical isolates. FEMS Microbiol Lett
481 270:265–271.
482 7. De Smet AMGA, Kluytmans JAJW, Cooper BS, Mascini EM, Benus RFJ, Van Der Werf TS,
483 Van Der Hoeven JG, Pickkers P, Bogaers-Hofman D, Van Der Meer NJM, Bernards AT,
484 Kuijper EJ, Joore JCA, Leverstein-van Hall MA, Bindels AJGH, Jansz AR, Wesselink RMJ, De

485 Jongh BM, Dennesen PJW, Van Asselt GJ, Velde LFT, Frenay IHME, Kaasjager K, Bosch FH,
486 Van Iterson M, Thijsen SFT, Kluge GH, Pauw W, De Vries JW, Kaan JA, Arends JP, Aarts
487 LPHJ, Sturm PDJ, Harinck HIJ, Voss A, Uijtendaal E V., Pharm D, Blok HEM, Groen EST,
488 Pouw ME, Kalkman CJ, Bonten MJM. 2009. Decontamination of the digestive tract and
489 oropharynx in ICU patients. N Engl J Med 360:20–31.
490 8. Aardema H, Bult W, Van Hateren K, Dieperink W, Touw DJ, Alffenaar JWC, Zijlstra JG.
491 2020. Continuous versus intermittent infusion of cefotaxime in critically ill patients: A
492 randomized controlled trial comparing plasma concentrations. J Antimicrob Chemother
493 75:441–448.
494 9. Wake DB. 1991. Homoplasy: the result of natural selection, or evidence of design
495 limitations? Am Nat 138:543–567.
496 10. Crispell J, Balaz D, Gordon SV. 2019. Homoplasyfinder: A simple tool to identify
497 homoplasies on a phylogeny. Microb Genomics 5.
498 11. Kluge AG, Farris JS. 1969. Quantitative phyletics and the evolution of anurans. Syst Biol
499 18:1–32.
500 12. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher
501 EM, Calver A, Sloutsky A, Kaur D, Posey JE, Plikaytis B, Oggioni MR, Gardy JL, Johnston JC,
502 Rodrigues M, Tang PKC, Kato-Maeda M, Borowsky ML, Muddukrishna B, Kreiswirth BN,
503 Kurepina N, Galagan J, Gagneux S, Birren B, Rubin EJ, Lander ES, Sabeti PC, Murray M.
504 2013. Genomic analysis identifies targets of convergent positive selection in drug-
505 resistant Mycobacterium tuberculosis. Nat Genet 45:1183–1189.
506 13. Mortimer TD, Weber AM, Pepperell CS. 2018. Signatures of Selection at Drug Resistance

507 Loci in Mycobacterium tuberculosis. mSystems 3:1–9.
508 14. Ruesen C, Chaidir L, van Laarhoven A, Dian S, Ganiem AR, Nebenzahl-Guimaraes H,
509 Huynen MA, Alisjahbana B, Dutilh BE, van Crevel R. 2018. Large-scale genomic analysis
510 shows association between homoplastic genetic variation in Mycobacterium tuberculosis
511 genes and meningeal or pulmonary tuberculosis. BMC Genomics 19.
512 15. Miotto P, Cabibbe AM, Feuerriegel S, Casali N, Drobniewski F, Rodionova Y, Bakonyte D,
513 Stakenas P, Pimkina E, Augustynowicz-Kopeć E, Degano M, Ambrosi A, Hoffner S, Mansjö
514 M, Werngren J, Rüsch-Gerdes S, Niemann S, Cirillo DM. 2014. Mycobacterium
515 tuberculosis pyrazinamide resistance determinants: A multicenter study. MBio 5:1–10.
516 16. Coolen JPM, Den Drijver EPM, Kluytmans JAJW, Verweij JJ, Lamberts BA, Soer JACJ,
517 Verhulst C, Wertheim HFL, Kolwijck E. 2019. Development of an algorithm to discriminate
518 between plasmid- and chromosomal-mediated AmpC β-lactamase production in
519 Escherichia coli by elaborate phenotypic and genotypic characterization. J Antimicrob
520 Chemother 74:3481–3488.
521 17. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood
522 trees for large alignments. PLoS One 5:e9490.
523 18. Crispell J, Benton CH, Balaz D, De Maio N, Ahkmetova A, Allen A, Biek R, Presho EL, Dale J,
524 Hewinson G, Lycett SJ, Nunez-Garcia J, Skuce RA, Trewby H, Wilson DJ, Zadoks RN,
525 Delahay RJ, Kao RR. 2019. Combining genomics and epidemiology to analyse bi-
526 directional transmission of mycobacterium bovis in a multi-host system. Elife 8:1–36.
527 19. Van Dorp L, Gelabert P, Rieux A, De Manuel M, De-Dios T, Gopalakrishnan S, Carøe C,
528 Sandoval-Velasco M, Fregel R, Olalde I, Escosa R, Aranda C, Huijben S, Mueller I,

529 Marquès-Bonet T, Balloux F, Gilbert MTP, Lalueza-Fox C, Shapiro B. 2020. Plasmodium
530 vivax Malaria Viewed through the Lens of an Eradicated European Strain. Mol Biol Evol
531 37:773–785.
532 20. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR.
533 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole
534 genome sequences using Gubbins. Nucleic Acids Res 43:e15.
535 21. Nelson EC, Gay Elisha B. 1999. Molecular basis of ampC hyperproduction in clinical
536 isolates of Escherichia coliAntimicrobial Agents and Chemotherapy.
537 22. Caroff N, Espaze E, Gautreau D, Richet H, Reynaud A. 2000. Analysis of the effects of -42
538 and -32 ampC promoter mutations in clinical isolates of Escherichia coli hyperproducing
539 AmpC. J Antimicrob Chemother 45:783–788.
540 23. Mulvey MR, Bryce E, Boyd DA, Ofner-Agostini M, Land AM, Simor AE, Paton S. 2005.
541 Molecular characterization of cefoxitin-resistant Escherichia coli from Canadian hospitals.
542 Antimicrob Agents Chemother 49:358–365.
543 24. Martínez-Martínez L. 2008. Extended-spectrum β-lactamases and the permeability
544 barrier. Clin Microbiol Infect 14:82–89.
545 25. Nordmann P, Mammeri H. 2007. Extended-spectrum cephalosporinases: Structure,
546 detection and epidemiology. Future Microbiol 2:297–307.
547 26. Yona AH, Alm EJ, Gore J. 2018. Random sequences rapidly evolve into de novo
548 promoters. Nat Commun 9:1530.
549 27. Liu S, Libchaber A. 2006. Some aspects of E. coli promoter evolution observed in a
550 molecular evolution experiment. J Mol Evol 62:536–550.

551 28. Korotkov K V., Sandkvist M, Hol WGJ. 2012. The type II secretion system: Biogenesis,
552 molecular architecture and mechanism. Nat Rev Microbiol.
553 29. Ho TD, Davis BM, Ritchie JM, Waldor MK. 2008. Type 2 secretion promotes
554 enterohemorrhagic Escherichia coli adherence and intestinal colonization. Infect Immun
555 76:1858–1865.
556 30. Baldi DL, Higginson EE, Hocking DM, Praszkier J, Cavaliere R, James CE, Bennett-Wood V,
557 Azzopardi KI, Turnbull L, Lithgow T, Robins-Browne RM, Whitchurch CB, Tauschek M.
558 2012. The type II secretion system and its ubiquitous lipoprotein substrate, SslE, are
559 required for biofilm formation and virulence of enteropathogenic Escherichia coli. Infect
560 Immun 80:2042–2052.
561 31. Kulkarni R, Dhakal BK, Slechta ES, Kurtz Z, Mulvey MA, Thanassi DG. 2009. Roles of
562 putative type II secretion and type IV pilus systems in the virulence of uropathogenic
563 Escherichia coli. PLoS One 4.
564 32. Dunne KA, Chaudhuri RR, Rossiter AE, Beriotto I, Browning DF, Squire D, Cunningham AF,
565 Cole JA, Loman N, Henderson IR. 2017. Sequencing a piece of history: Complete genome
566 sequence of the original Escherichia coli strain. Microb Genomics 3.
567 33. Vanderlinde EM, Strozen TG, Hernández SB, Cava F, Howard SP. 2017. Alterations in
568 peptidoglycan cross-linking suppress the secretin assembly defect caused by mutation of
569 GspA in the type II secretion system. J Bacteriol 199.
570 34. Juan C, Torrens G, Barceló IM, Oliver A. 2018. Interplay between Peptidoglycan Biology
571 and Virulence in Gram-Negative Pathogens. Microbiol Mol Biol Rev 82.
572 35. Read TD, Massey RC. 2014. Characterizing the genetic basis of bacterial phenotypes using

573 genome-wide association studies: A new direction for bacteriology. Genome Med.
574 BioMed Central Ltd.
575 36. Chen PE, Shapiro BJ. 2015. The advent of genome-wide association studies for bacteria.
576 Curr Opin Microbiol. Elsevier Ltd.
577 37. Shapiro BJ, David LA, Friedman J, Alm EJ. 2009. Looking for Darwin's footprints in the
578 microbial world. Trends Microbiol 17:196–204.
579 38. Benjak A, Avanzi C, Singh P, Loiseau C, Girma S, Busso P, Fontes ANB, Miyamoto Y,
580 Namisato M, Bobosha K, Salgado CG, Da Silva MB, Bouth RC, Frade MAC, Filho FB,
581 Barreto JG, Nery JAC, Bührer-Sékula S, Lupien A, Al-Samie AR, Al-Qubati Y, Alkubati AS,
582 Bretzel G, Vera-Cabrera L, Sakho F, Johnson CR, Kodio M, Fomba A, Sow SO, Gado M,
583 Konaté O, Stefani MMA, Penna GO, Suffys PN, Sarno EN, Moraes MO, Rosa PS, Baptista
584 IMFDi, Spencer JS, Aseffa A, Matsuoka M, Kai M, Cole ST. 2018. Phylogenomics and
585 antimicrobial resistance of the leprosy bacillus Mycobacterium leprae. Nat Commun 9.
586 39. Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous
587 mutations in the bacterium Escherichia coli as determined by whole-genome sequencing.
588 Proc Natl Acad Sci U S A 109.
589 40. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of
590 commensal Escherichia coli. Nat Rev Microbiol 8:207–217.
591 41. Escobar-Páramo P, Sabbagh A, Darlu P, Pradillon O, Vaury C, Denamur E, Lecointre G.
592 2004. Decreasing the effects of horizontal gene transfer on bacterial phylogeny: The
593 Escherichia coli case study. Mol Phylogenet Evol 30:243–250.
594 42. Price LB, Johnson JR, Aziz M, Clabots C, Johnston B, Tchesnokova V, Nordstrom L, Billig M,

595 Chattopadhyay S, Stegger M, Andersen PS, Pearson T, Riddell K, Rogers P, Scholes D, Kahl
596 B, Keim P, Sokurenko E V. 2013. The epidemic of extended-spectrum-β-lactamase-
597 producing Escherichia coli ST131 is driven by a single highly pathogenic subclone, H30-Rx.
598 MBio 4:1–10.
599 43. Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is
600 robust to recombination but demographic inference is not. MBio 5.
601 44. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko
602 SI, Pham S, Prjibelski AD, Pyshkin A V., Sirotkin A V., Vyahhi N, Tesler G, Alekseyev MA,
603 Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to
604 single-cell sequencing. J Comput Biol 19:455–477.
605 45. Beghain J, Bridier-Nahmias A, Nagard H Le, Denamur E, Clermont O. 2018.
606 ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain
607 phylotyping. Microb Genomics 4:1–8.
608 46. Seemann T. mlst. Github https://github.com/tseemann/mlst.
609 47. Jolley KA, Maiden MCJ. 2010. BIGSdb: Scalable analysis of bacterial genome variation at
610 the population level. BMC Bioinformatics 11.
611 48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search
612 tool. J Mol Biol 215:403–410.
613 49. Peter-Getzlaff S, Polsfuss S, Poledica M, Hombach M, Giger J, Böttger EC, Zbinden R,
614 Bloemberg G V. 2011. Detection of AmpC beta-lactamase in Escherichia coli: Comparison
615 of three phenotypic confirmation assays and genetic analysis. J Clin Microbiol 49:2924–
616 2932.

617 50. Seemann T. Abricate. Github.

618 51. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Unicycler: Resolving bacterial genome
619 assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595.

620 52. Tatusova T, Dicuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze
621 A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline.
622 Nucleic Acids Res 44:6614–6624.

623 53. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F,
624 Derbyshire MK, Gonzales NR, Gwadz M, Lu F, Marchler GH, Song JS, Thanki N, Yamashita
625 RA, Zheng C, Thibaud-Nissen F, Geer LY, Marchler-Bauer A, Pruitt KD. 2018. RefSeq: An
626 update on prokaryotic genome annotation and curation. Nucleic Acids Res 46:D851–
627 D860.

628 54. European Committee on Antimicrobial Susceptibility Testing (EUCAST). 2020. The
629 European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for
630 interpretation of MICs and zone diameters. Version 10.0, 2020.

631 55. Mehta CR, Patel NR. 1983. A network algorithm for performing fisher's exact test in r × c
632 contingency tables. J Am Stat Assoc 78:427–434.

633 56. Benjamini Y, Hochberg Y. 1995. <Benjamini&Hochberg1995_FDR.pdf>. J R Stat Soc Ser B
634 57:289–300.

635 57. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.
636 2012. A program for annotating and predicting the effects of single nucleotide
637 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;
638 iso-2; iso-3. Fly (Austin) 6:80–92.

639 58. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. 2020. The EnteroBase user's guide,
640 with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia
641 core genomic diversity. Genome Res 30:138–152.
642 59. Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: Recent updates and new
643 developments. Nucleic Acids Res 47:W256–W259.
644 60. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo
645 generator. Genome Res 14:1188–1190.
646 61. Connors J, Krzywinski M, Schein J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009.
647 Circos : An information aesthetic for comparative genomics. Genome Res 19:1639–1645.
648 62. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. 2018.
649 Phandango: An interactive viewer for bacterial population genomics. Bioinformatics
650 34:292–293.
651

652 **CAPTIONS**

653 **FIG 1** Schematic of workflow used to perform the homoplasy-based association analysis.
654 Starting from the top A) the *de novo* assembly of the NextSeq/MiSeq reads and B) the hybrid
655 assembly of the reference chromosome ampC_069. On the left side C) the alignment of
656 promoter/attenuator region. In the middle D) the coreSNP analysis for the phylogeny used in E)
657 the homoplasy analysis combined with F) the fullSNP data on the right, which was also used for
658 G) the statistics (Fisher Exact & FDR) to relate cefotaxime (CTX) resistance to SNP positions. H)
659 Inferring recombination events using Gubbins.
660

661 **FIG 2** Approximately maximum-likelihood phylogenetic tree of all 172 *E. coli* isolates based on
662 the coreSNP alignment with the resistance for cefotaxime (CTX), *pampC* gene presence, MLSTs,
663 phylogroups, and the alignments of the promoter and the attenuator region. Positions with a
664 CAT likelihood score ≤60% are indicated as red dots.
665
666 **FIG 3** Sequence logo with probability score for the promoter and the attenuator region. The
667 Consistency Index and the minimum number of changes on the tree per position are
668 represented below the sequence logos.
669
670 **FIG 4** Circos plot for the full chromosome of ampC_0069 (accession no. CP046396) with per
671 position the various metrics used. A) The blue colored ring represents the Consistency Index
672 results per genomic position. The two red dots indicate the -42 and -32 position on the
673 promoter. The black circle line indicates the 0.07142857 Consistency Index value. B) The ring
674 with a red background shows all positions that were significantly associated to cefotaxime (CTX)
675 resistance in all non-*pampC* harboring *E. coli* isolates. Larger bars pointing outwards indicate
676 multiple significant associated positions in a small genomic region. C) The ring with the green
677 background shows all 24 positions that have a low Consistency Index of ≤ 0.05882353 and are
678 significantly associated with CTX resistance in all non-*pampC* harboring *E. coli* isolates.
679
680 **TABLE 1** Table of the distribution of AmpC promoter and attenuator variants as well as
681 the amount of different MLST and phylogroups per grouped genotype (*pampC*, hyperproducers
682 and low-level AmpC producers).

683

684 **TABLE 2** The *n*=24 positions with a significant association with cefotaxime resistance

685 (FDR ≤0.05) and with a consistency index ≤ 0.05882353.

686

687 **FIG S1** Violin plots of the log10 Consistency Indexes of the promoter and attenuator.

688

689 **FIG S2** Distribution of the log10 Consistency Indexes of all genomic position based on the *E. coli*

690 ampC_0069 reference chromosome, compared to the log10 Consistency Indexes of the

691 promoter and attenuator region.

692

693 **FIG S3** Recombination events inferred from all 172 *E. coli* isolates by Gubbins displayed along

694 the approximately maximum-likelihood phylogenetic tree based on the coreSNP alignment.

695 Phylogroups are depicted as in FIG 2. Gubbins blocks are colored red if they are ancestral, and

696 blue if they only affect one isolate. The line graph represents the recombination prevalence

697 along the sequence. The 24 positions with a significant association with cefotaxime resistance

698 (FDR ≤0.05) and a consistency index ≤ 0.05882353 are indicated on the top of the figure. The

699 two missense mutations and *ampC* promoter region are displayed in blue.

700

701 **Table S1** Classification of n=172 *E. coli* isolates in the three genotypes (*pampC*,

702 hyperproducer, low-level AmpC producers), with the results of the MLST and phylogroups

703 stratification and the different mutations in the promoter and attenuator per isolate. Isolates

704 with a CTX MIC >2 mg/L without a confirmed *pampC* gene are depicted in **bold**.

705

706 **Table S2** SNP analysis for n=172 *E. coli* isolates according to snippy statistics.

707

708

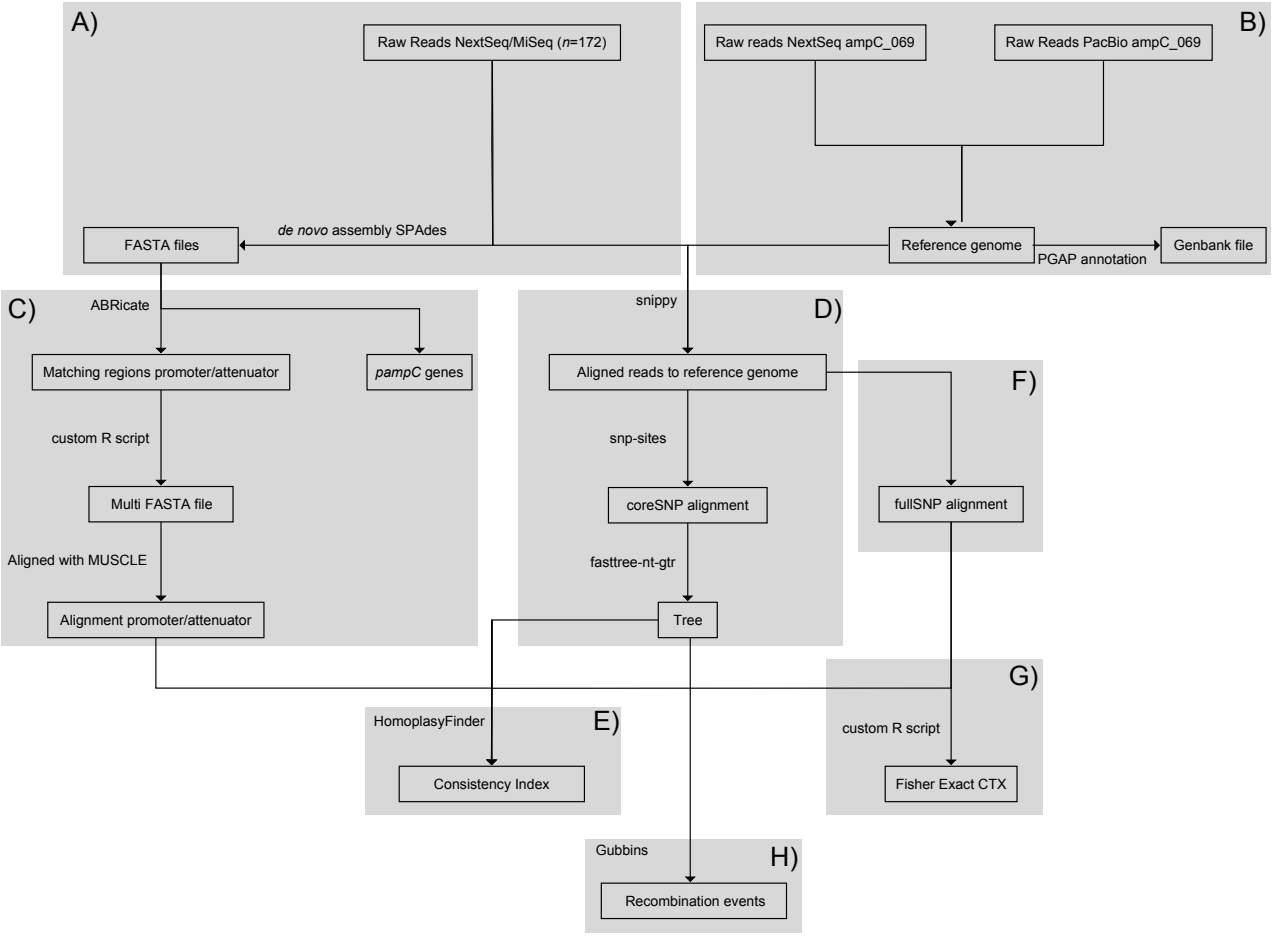

FIG 1 Schematic of workflow used to perform the homoplasy-based association analysis. Starting from the top A) the *de novo* assembly of the NextSeq/MiSeq reads and B) the hybrid assembly of the reference chromosome ampC_069. On the left side C) the alignment of promoter/attenuator region. In the middle 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.

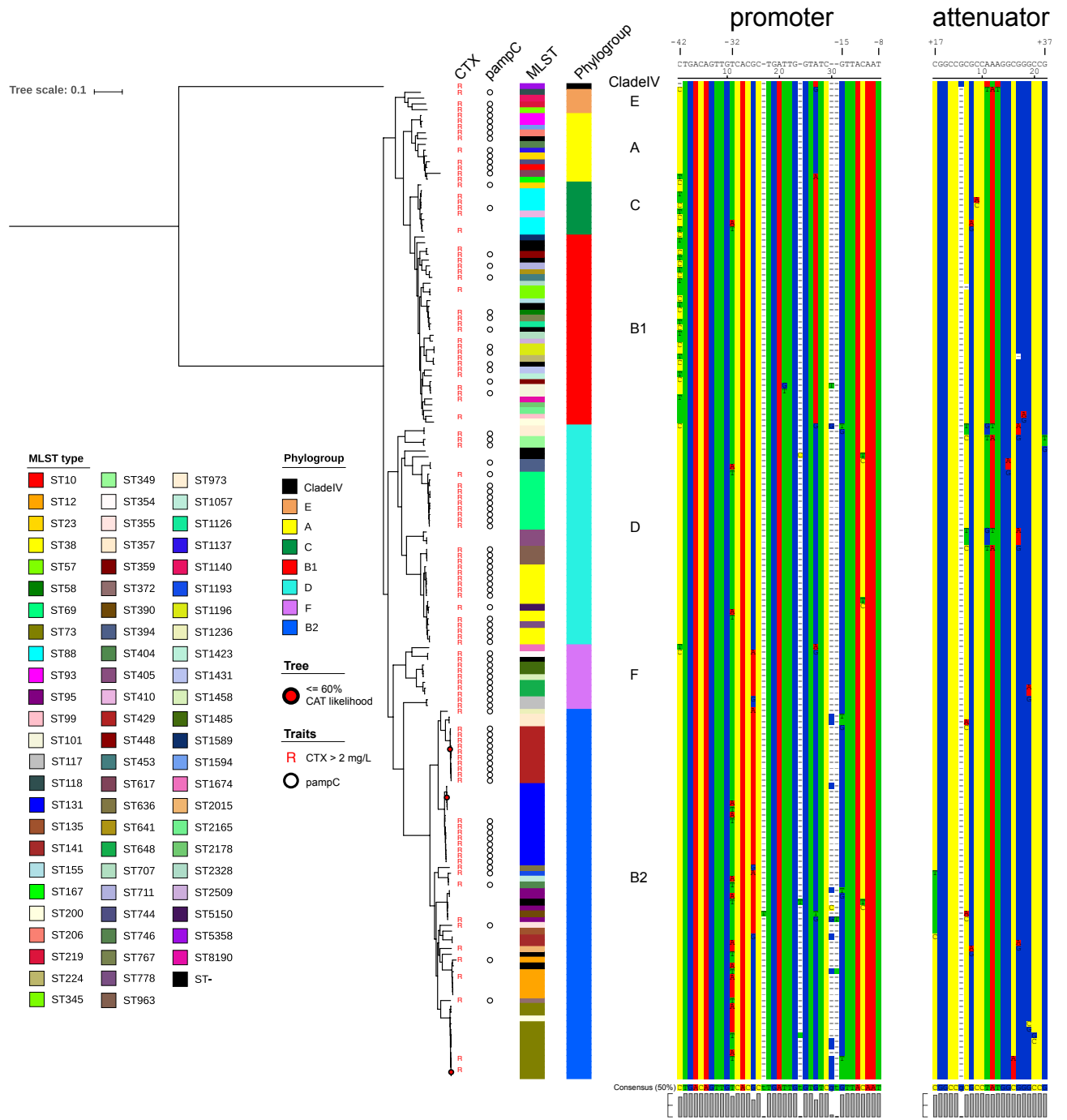

FIG 2 Approximately maximum-likelihood phylogenetic tree of all 172 *E. coli* isolates based on the coreSNP alignment with the resistance for cefotaxime (CTX), *pampC* gene presence, MLSTs, phylogroups, and the alignments of the promoter and the attenuator region. Positions with a CAT likelihood score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

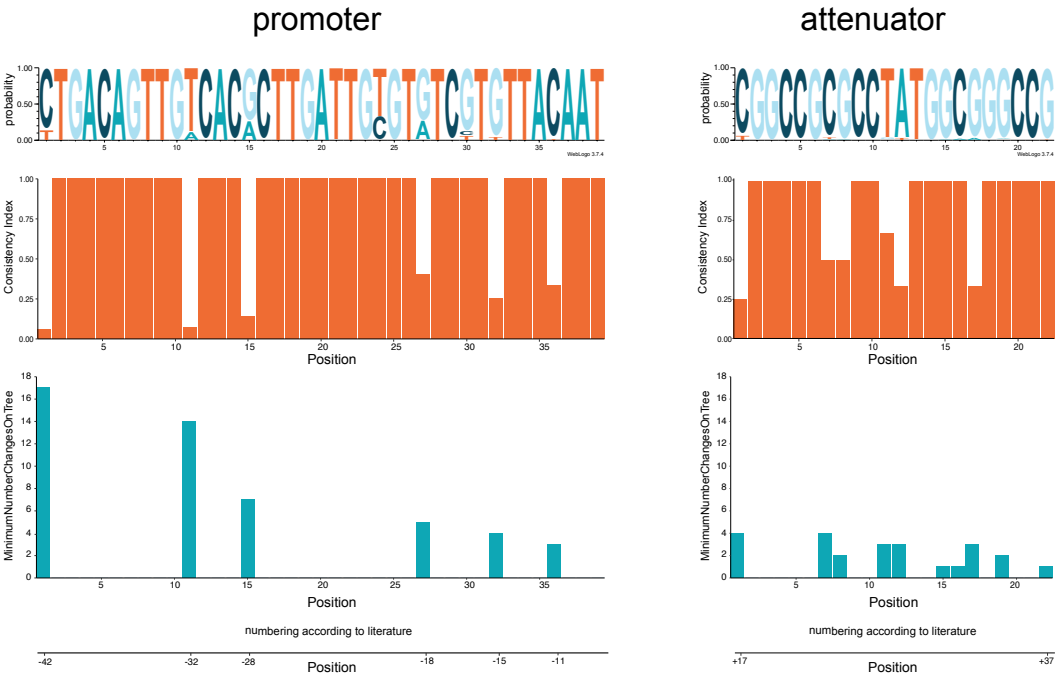

FIG 3 Sequencelogo with probability score for the promoter and the attenuator region. The Consistency Index and the minimum number of changes on the tree per position are represented below the sequence logos.

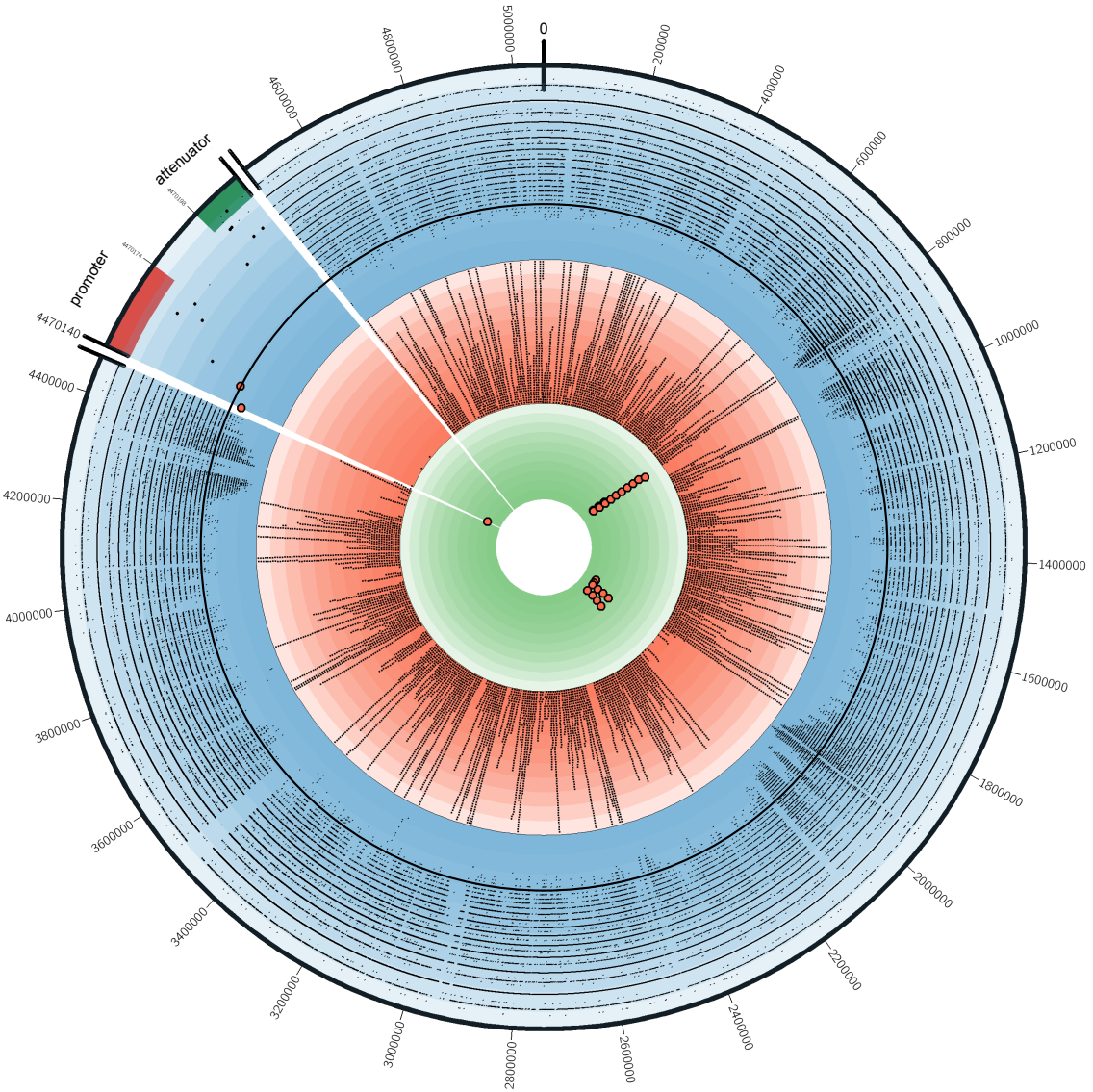

FIG 4 Circos plot for the full chromosome of ampC_0069 (accession no. CP046396) with per position the various metrics used. A) The blue colored ring represents the Consistency Index results per genomic position. The two red dots indicate the -42 and -32 position on the promoter. The black circle line 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.

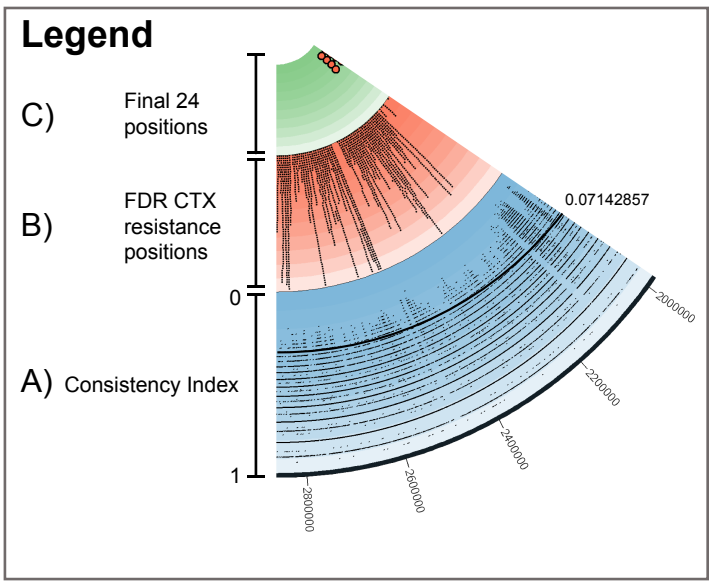

TABLE 1

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