The population genomics of increased virulence and antibiotic resistance in human commensal *Escherichia coli* over 30 years in France

Julie Marin[a,b#], Olivier Clermont[a,b], Guilhem Royer[a,b,c,d], Mélanie Mercier-Darty[e], Jean Winoc Decousser[d,f], Olivier Tenaillon[a,b], Erick Denamur[a,b,g*] and François Blanquart[a,b,h*#].

[a]Université Sorbonne Paris Nord, INSERM, IAME, 93017 Bobigny, France

[b]Université de Paris Cité, INSERM, IAME, 75018 Paris, France

[c]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91042 Evry, France

[d]AP-HP, Département de prévention, diagnostic et traitement des infections, Hôpital Henri Mondor, 94000 Créteil, France

[e]AP-HP, Next Generation Sequencing Platform, University Hospital Henri Mondor, 94000 Créteil, France

[f]EA 7380 Dynamyc Univ Paris Est Creteil (UPEC), Ecole Nationale Vétérinaire d'Alfort (EnvA), Faculté de Médecine de Créteil, 94000 Créteil, France

[g]AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, 75018 Paris, France

[h]Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, 75005 Paris, France

***: co-last authors**

ORCID numbers: 0000-0002-7160-9302 (JM), 0000-0001-9092-8336 (GR), 0000-0001-7105-6202 (JWD), 0000-0002-3796-1601 (OT), 0000-0002-7860-1717 (ED), 0000-0003-0591-2466 (FB)

Running title: Virulence and resistance evolution of *E. coli*

Corresponding authors: francois.blanquart@college-de-france.fr, julie.marin@univ-paris13.fr

ABSTRACT

Escherichia coli is a commensal species of the lower intestine, but also a major pathogen causing intestinal and extra-intestinal infections, increasingly prevalent and resistant to antibiotics. Most studies on genomic evolution of *E. coli* used isolates from infections. Here instead, we whole-genome sequenced a collection of 403 commensal *E. coli* isolated from fecal samples of healthy adult volunteers in France (1980-2010). These isolates were distributed mainly in phylogroups A and B2 (30% each) and belonged among 152 sequence types (STs), the five most frequent being ST10 (16.3%), ST73 (6.3%), ST95 (5.0%), ST69 (4.2%) and ST59 (3.9%), and 224 O:H serotypes. ST and serotype diversity increased over time. The O1, O2, O6 and O25-groups used in bioconjugate O-antigen vaccine against extraintestinal infections were found in 23% of the strains of our collection. Extra-intestinal virulence-associated genes and antibiotic resistances increased in frequency, suggesting positive selection on these genes. Evolution of virulence genes was driven by both clonal expansion of STs with more virulence genes ("ST-driven") and increases in gene frequency within STs independently of changes in ST frequencies ("gene-driven"), whereas the evolution of resistance was dominated by increases in frequency within STs ("gene-driven"). This study provides a unique picture of the phylogenomic evolution of *E. coli* in its human commensal habitat over 30 years and will have implications for the development of preventive strategies.

IMPORTANCE

Escherichia coli is an opportunistic pathogen with the greatest burden of antibiotic resistance, one of the main causes of nosocomial infections and an increasing concern in an ageing population. Deciphering the evolutionary dynamics of virulence and antibiotic resistance in commensal *E. coli* is important to understand adaptation and anticipate future changes. The gut of vertebrates is the primary habitat of *E. coli* and probably where selection for virulence and resistance take place.

Unfortunately, most whole-genome sequenced strains are isolated from pathogenic conditions. Here instead, we whole genome sequenced 403 *E. coli* commensals isolated from healthy French subjects on a 30-year period. Virulence and resistance genes increased in frequency both by clonal expansion of clones carrying them and by increases in frequency within clones. Such studies should be performed worldwide to have a broader picture of *E. coli* evolution and adaptation in commensalism.

Keywords: antibiotic resistance, commensal, *Escherichia coli*, evolution, genomics, virulence.

61 **INTRODUCTION**

62 *Escherichia coli* is a commensal species of the lower intestine of humans and other vertebrates (1).
63 It is also a major opportunistic pathogen causing intestinal and extra-intestinal infections (2),
64 ranked third in the World Health Organization list of antibiotic resistant 'priority pathogens', as it
65 is today the bacterial pathogen associated with the largest burden of antimicrobial resistance (3).

66 *E. coli* strains share a core genome of about 2,000 genes over 4,700 genes in average (4).
67 In spite of numerous recombination events, this species presents a robust clonal structure with at
68 least nine phylogroups called A, B1, B2, C, D, E, F, G and H (2), and sequence types (STs) defined
69 at a finer genetic resolution. In addition, molecular typing of genes coding for surface structures as
70 O-polysaccharide and H-flagellar antigens (5), as well as the FimH protein (6), allows, in
71 conjunction with STs, a better characterization of the clones (7). O-antigens are key elements in
72 the pathophysiological processes and used as a target for the development of *E. coli* vaccines to
73 protect against extra-intestinal infections (8, 9).

74 From a public health perspective, it is particularly important to understand the evolution of
75 the repertoire of virulence and antibiotic resistance genes. Virulence genes are associated with a
76 higher risk of urinary tract and bloodstream infections (10, 11). Antibiotic resistant infections are
77 associated with longer hospital stay, increased risk of death and public health costs (12). Both
78 virulence and antibiotic resistance genes repertoire greatly varies among strains, with continuous
79 gene acquisition and loss (13). Temporal trends in virulence have been less often examined than
80 trends in resistance. The mean number of virulence genes in commensal *E. coli* has increased from
81 1980 to 2010 in France (14). Increased virulence of *E. coli* could be one explanation for the
82 increasing incidence of *E. coli* bacteremia in Europe (12, 15), together with increased reporting or
83 changing epidemiological factors like the ageing population. Resistances to multiple antibiotics

84 have concomitantly rapidly increased in frequency in *E. coli* over the last decades, and stabilized
85 at an intermediate level (16).

86 What evolutionary forces act on virulence and resistance genes? A first hypothesis is that
87 these genes are direct targets of selection. Virulence genes are thought to primarily be selected in
88 the intestine as a by-product of commensalism (17, 18). Virulence genes are linked with longer
89 persistence in the gut (19). Extra-intestinal compartments do not typically lead to efficient onward
90 transmission and are often considered an evolutionary dead end (18). Alternatively, clonal
91 expansion could modify the prevalence of resistance and virulence genes, without these genes
92 being the direct target of selection. Clonal expansion may be important in *E. coli* since several
93 resistance and virulence genes are closely linked to specific STs (20, 21). The selective forces
94 acting on resistance are clearer. Resistance is selected by antibiotic use, as evidenced by the spatial
95 correlation between local antibiotic use and resistance (22), and the fact that hosts are colonized by
96 resistant strains more frequently after antibiotic use (23). For *E. coli*, exposure to antibiotics results
97 95% of the time from antibiotic treatment prescribed for infections unrelated to *E. coli*, and not to
98 treat an *E. coli* infection (24). Such "bystander" antibiotic exposure happens around once per year
99 in adults in France (25).

100 Although the gut is the typical habitat of *E. coli* and likely the main ecological context of
101 selection for virulence and resistance, most studies on the evolution of virulence and antibiotic
102 resistance focused on pathogenic clinical strains isolated from extra-intestinal infections (26–28).
103 Extra-intestinal infections are a rare occurrence in *E. coli*'s life: the incidence of urinary tract
104 infections is of the order of 10^{-2} per person-year (29), and of bacteremia of the order of 10^{-4} per
105 person-year (30). In comparison to commensal *E. coli*, *E. coli* sampled from infections are biased
106 towards strains carrying more virulent genes (11, 26, 27). Strains from infections may also be less
107 diverse than commensal strains as they belong to some specific clones (2, 31), hindering the

108 detection of temporal trends in these elements and the evolutionary mechanisms driving the
109 emergence of pathogenic clones (32).

110 To avoid these biases, we investigated here the short-term genomic evolution of commensal
111 *E. coli* over 1980 to 2010 using a rare collection of 403 strains from fecal samples of healthy
112 volunteers in France. We assembled this collection from five previously described collections all
113 collected in healthy individuals with a diversity of ages, sex, and located in two French regions
114 (table S1), and controlled for these covariates in all analyses.

115

116 **RESULTS**

117 The pan-genome of the 403 strains contained 47,367 genes, with no saturation in the number of
118 genes detected (figure S1). The core genome, defined as the number of genes found in more than
119 99% of the isolates, contained 2156 genes. A total of 504 genes were found in 95% to 99% of the
120 isolates and 3252 genes were found in 15% to 95% of the isolates. The large majority of genes
121 (41,455 out of 47,367) were found in less than 15% of the isolates.

122

123 ***1. Evolution of phylogenetic groups, STs and serotypes over time***

124 Of the 403 isolates, 398 were assigned to the *E. coli* phylogroups A (120, 30%), B1 (58, 14%),
125 B2 (120, 30%), C (10, 2%), D (44, 11%), E (13, 3%), F (26, 6%), G (6, 1%) and H (1, 0.2%) (tables
126 S2-3). We found two strains (0.5%) belonging to the *Escherichia* clade IV and one (0.2%) to the
127 *Escherichia* clade V. Two strains (C020-013 and LBC20a, 0.5%) were phylogenetically distinct
128 from the known phylogroups (A to H). They probably belong to an undescribed phylogroup (figure
129 S2). We fitted a binomial generalized linear model (GLM) for the presence/absence of each
130 phylogroup as a function of time, controlling for host age, region, sex. The use of antibiotics in the
131 past month was an exclusion criterion in four of the five studies generating the collections. For the

132 last collection, we controlled for the possible (but unlikely) use of antibiotic by adding a specific
133 factor for this collection in the regression (table S1). We detected a significant decrease in
134 frequency for the phylogroup A (from 58% in 1980 to 26% in 2010, p-value=1.79e-03) and a
135 significant increase in frequency for the phylogroup B2 (from 9% in 1980 to 36% in 2010, p-
136 value=1.34e-03) (figure 1A). No significant change was detected for the other phylogroups.

137 Those 403 isolates were distributed among 152 Warwick University scheme STs (33). The five
138 most frequent STs were ST10 (15.4%) (phylogroup A), followed by ST73 (5.9%) and ST95 (4.7%)
139 (both of phylogroup B2), ST69 (4.0%) (phylogroup D) and ST59 (3.7%) (phylogroup F) (figure
140 1B and table 1). Using a binomial GLM, we detected a significant decrease in frequency for ST10
141 between 1980 and 2010 (from 29% in 1980 to 15% in 2010, p-value=0.028). No other significant
142 change was detected for the three others most frequent STs. A total of 30, 59 and 98 distinct STs
143 were detected in 1980, 2001 (2000-2002) and 2010 respectively. ST diversity increased over time
144 even after accounting for the increasing number of strains sampled, and the changes in phylogroup
145 frequencies (figures 1B and S3).

146 We next typed the isolates using the diversity of genes coding for O and H antigens and FimH
147 protein. These genes are located in the two major hot spots of recombination along the *E. coli*
148 chromosome (4) and the corresponding antigens are important for the host immune response and
149 pathogenesis. Serotype (O:H combination) diversity increased between 1980 and 2010, however,
150 the diversity of O-groups, H-types and *fimH* alleles remained stable between 2001 and 2010 when
151 considered individually (figure S4 and tables S4-7). The four O-groups targeted by the recently
152 developed bioconjugate vaccine ExPEC4V (8, 9), O1, O2, O6 and O25 represented 24% of our
153 collection. By fitting a binomial generalized linear model (GLM), we did not detect a change in
154 frequency between 1980 and 2010 (figure 1C).

155 In the analyses of temporal trends in phylogenetic groups, STs, and serotypes over time, we
156 controlled for the covariates host age, regions, sex and possible use of antibiotics. In this series of
157 linear models, only 3 out of 88 inferred effects of control covariates were significant at the 0.05
158 level. This suggests these covariates have no or negligible effect on the genetic composition of *E.*
159 *coli*.

160 We next examined how the diversity of O and H antigen and FimH protein coding genes
161 evolves in relation to the phylogenetic structure of three STs that we were able to date: ST10, ST69
162 and ST95. ST10 strains exhibit a large diversity of O and H antigens (38 serotypes) and *fimH* alleles
163 (15 alleles). ST69 strains exhibit a limited diversity of O serogroups (O15, O17, O25 and O45)
164 associated with H18 and *fimH*27 almost exclusively (figure 2). We found six serotypes and five
165 *fimH* alleles among ST95 strains. The apparent diversity in O:H combinations and *fimH* alleles
166 reflects in part the age of the ST that can vary by a 20-fold factor according to our estimations
167 (figure 2). The MRCA (most recent common ancestor) age of ST10, 1377 [866-1637], is much
168 older than the MRCA age of ST69, 1951 [1924-1969], and ST95, 1680 [766-1882] (Bayesian
169 skyline model, figure S5 and table S8). We tested whether the difference in their timescale of
170 evolution explains the difference in diversity and found that the diversity of genes encoding for O
171 and H antigens and for the FimH protein increased faster for ST69 than for ST10 and ST95 (figure
172 S6). We also quantified the change in diversity of epidemiological types (O:H serotype and *fimH*
173 allele) within the most abundant ST of our commensal collection (ST10). We did not detect a
174 significant change in frequency for any of the five most frequent serotypes within ST10 between
175 1980 and 2010 (figure S7).

176

177 ***2. Recent temporal evolution of virulence and resistance in commensal* E. coli**

178 We first evaluated the effect of phylogroup frequency changes and time on the recent temporal
179 evolution of virulence and resistance. Next, we analyzed the gene frequency variations over two
180 periods of time, between 1980 and 2001 and between 2001 and 2010. We used Fisher's exact test
181 (gene presence/absence *versus* date as a category) to test the temporal changes in gene frequency
182 over each time period. We then evaluated the action of directional selection on gene presence with
183 a logistic model (gene presence/absence *versus* time as a continuous variable) between 1980 and
184 2010. Finally, we computed dN/dS ratios to test if gene sequences evolved neutrally or were under
185 diversifying or purifying selection.
186
187 *a. Recent increase in virulence and resistance*
188 The correspondence analysis indicated a significant association between the phylogroups and the
189 virulence and resistance categories (chi-squared test, p-value < 2.2e-16) (figure S8 and tables S9-
190 11). The phylogroup B2 carry more VFs distributed in the adhesin, invasin, iron acquisition and
191 protectin categories, but fewer resistance genes. The phylogroups A, B1 and E carry more toxin
192 genes. The phylogroups C, D and the undescribed phylogroup are positively correlated with the
193 large majority of resistance categories, including resistances to beta-lactam, sulphonamide,
194 aminoglycoside, phenicol, tetracycline and trimethoprim antibiotics.
195 The prevalence of virulence increased between 1980 and 2010 (figure 3). First, the mean
196 virulence score (number of virulence factors out of the 101 assayed) increased from 14 to 19. Using
197 a linear model with sampling dates and phylogroups as predictors, we showed that, between 1980
198 and 2010, the increase in virulence score is explained by the change in the phylogroup frequency;
199 the effect of the sampling date is not significant (effect size = -1.90e-02 [-0.140; 0.102] VF product
200 per year, p-value = 0.743). This increase is driven by the change in frequency of several
201 phylogroups carrying many VFs, such as B2 (9% of B2 strains in 1980 to 36% in 2010 with a mean

202 virulence score of 24, effect size = +11 [6; 17] compared to phylogroup A, p-value = 7.37-04) and
203 F (4% of F strains in 1980 to 8% in 2010 with a mean virulence score of 23, effect size = +14 [7;
204 18] compared to phylogroup A, p-value = 1.87e-04). When testing each time period individually,
205 we did not find a significant effect of the sampling date either, but only of the phylogroup frequency
206 (between 1980 and 2001: effect size = +0.040 [-0.087; 0.167] VF product per year, p-value = 0.472;
207 between 2001 and 2010: effect size = -0.325 VF product per year [-0.765; 0.114], p-value = 0.126).
208 The frequency of resistance also increased through time. The fraction of strains resistant to
209 at least one antibiotic class increased from 19% to 44% from 1980 to 2010 (figure 3). This rise was
210 significant within a linear model with sampling dates and phylogroups as predictors (effect size =
211 +1.12e-02 [4.47e-03; 1.79e-02] per year, p-value = 2.88e-03), showing increased resistance
212 independently of changes in phylogroup frequencies. In addition to sampling dates, this increase is
213 also driven by the change in frequency of one minor phylogroup, phylogroup F which increases
214 from 4% in 1980 to 8% in 2010 and in which 88% of the strains are resistant to at least one antibiotic
215 (effect size = +0.680 [0.361; 1.00] compared to phylogroup A, p-value = 3.87e-04). When
216 considering each time period individually, the frequency of resistance to one antibiotic significantly
217 increased between 1980 and 2001 (effect size = +9.77e-03 [4.13e-04; 0.019] per year, p-value =
218 0.043). The increase between 2001 to 2010 was comparable in size but not significant (effect size
219 = +0.018 [-0.003; 0.039] per year, p-value = 0.082). Interestingly, the proportion of strains resistant
220 to two or more antibiotic categories remained stable at 27% between 2001 and 2010.
221
222 *b. ST clonal expansion versus increase in gene frequency within STs*
223 We next evaluated how the temporal variation in the frequency of individual resistance and
224 virulence genes was driven by clonal expansion or contraction of STs versus gene frequency
225 changes within one or several STs. We chose to do this analysis at the level of STs because they

226 are evolutionary units smaller than phylogroups but large enough to be followed through time, and
227 distinct in their gene content (34). Considering the size of our collection, we would lack power to
228 investigate temporal changes at a finer phylogenetic unit. Eight STs are present in 1980, 2001 and
229 2010, they belong to the phylogroups A, B2, D and F and represent 35% of our dataset (table S12).
230 We decomposed the temporal change in individual virulence and resistance gene frequencies in
231 two additive components (figures 4-6 and tables S13-16). First, "ST-driven" change corresponds
232 to a change in gene frequency driven by the change in frequency of STs (figure S9). Second, "gene-
233 driven" change reflects the change in gene frequency within STs, independently of changes in ST
234 frequencies.

235 The evolution of virulence is both ST and gene driven. The relative contribution of the two
236 processes depends on the time period considered. Frequency changes within STs (gene driven
237 change) contributed mostly in the period between 1980 and 2001 (figures 4A and 5A). Those
238 changes include notably the large increase in frequency of the *kps* operon (*kpsD*, *kpsE* and *kpsM*)
239 involved in the K1 capsule biosynthesis, from 40% to 78% (p-values = 0.002 for the association
240 between the gene presence and time of sample between 1980 and 2001 according to Fisher's exact
241 test), of the *iuc* operon (aerobactin, from 32-36% to 60-68%, p-values < 0.02), of the HPI (High
242 Pathogenicity Island, from 52% to 75%, p-values = 0.028 and 0.042) and of the *pap* operon
243 (pyelonephritis-associated pilus, from 12-24% to 25-64% , p-values < 0.02) (figure 4A and table
244 S13). From 2001 to 2010, the change in virulence was driven by both the increase in frequency of
245 more virulent STs (ST driven change), in particular those carrying genes coding for adhesins, and
246 the change in gene frequency within STs (gene driven change) (figure 4B).

247 In contrast to virulence, the evolution of resistance is almost exclusively gene driven and
248 occurred primarily between 1980 and 2001 (figures 5-6 and tables S15-16). Significant increase in
249 frequency was detected for genes responsible for resistance to beta-lactam ($bla_{\text{TEM-1B}}$, from 0% to

250 14%, p-value = 0.003), tetracycline (*tet*(A), from 0% to 14%, p-value = 0.004), sulphonamide (*sul*1,
251 from 0% to 14%, p-value = 0.005 and aminoglycoside (*aph*(6)-Id, from 0% to 18%, p-value = 0.02)
252 antibiotics. Resistance genes overall frequency did not change much in the second period between
253 2001 and 2010. We note however that any increasing gene frequency is ST-driven. Thus, these
254 genes rapidly increased in frequency in 20 years and subsequently stabilized, and their dynamics
255 were unaffected by changes in the ST composition of the population.

256 Gene content visualization of ST10, ST69 and ST95 corroborate these results (figure 2).
257 Virulence genes appear to be more phylogenetically clustered than resistance genes, explaining the
258 importance of clonal expansion in driving the evolution of the virulence gene repertoire. We last
259 tested if genes located on plasmids are more likely gene-driven (increasing simultaneously within
260 multiple STs) than ST-driven (clonal expansion), but found no significant effect (correlation across
261 genes between the proportion of gene driven change and plasmid *vs.* chromosome as a categorical
262 variable, effect size = + 0.09 [-0.02; 0.21], p-value = 9.62e-02).

263

264 *c. Further evidence for selection on resistance and virulence genes*

265 We detected a significant increase in frequency over at least one period of time (by Fisher's
266 exact test) for 59 virulence genes and four genes associated to antibiotic resistance. To test if the
267 presence of these genes (relative to absence) is under directional selection over the whole period
268 1980 to 2010, we fitted a logistic model for gene presence/absence as a function of time and
269 phylogroup for these 63 genes. In population genetics, the action of selection on a gene results in
270 a linear trend on the logit scale. Thus, the coefficient of a logistic model (binomial generalized
271 linear model) is interpretable as a selection coefficient (additionally controlling for association with
272 phylogroups). Out of the 18 genes with a significant coefficient (at the 0.05 level), 16 had a positive

273 coefficient and 2 a negative coefficient (table S17). This suggests that the presence of these genes
274 is under positive selection.

275 The two antibiotic resistance genes increasing in frequency for which we could compute
276 dN/dS (*tet*(A) and $bla_{\text{TEM-1B}}$), were under purifying selection (table S18). For virulence genes, we
277 found a significant difference in rates of synonymous and non-synonymous mutations for 37 genes
278 out of 55, with 30 under purifying selection (including *kps* operon and *papX*), and 7 under
279 diversifying selection (including the *iutA* transmembrane transporter involved in iron uptake) (table
280 S18).

281

282 **DISCUSSION**

283 ***1. The phylogenetic history of* E. coli *commensal strains and their serotypes***

284 The primary habitat of *E. coli* is the gut. However, most of the studies interested in virulence and
285 resistance to antibiotics focused on pathogenic collections isolated from extra-intestinal infections,
286 with a few exceptions (35, 32, 36). Here, using whole-genome analysis, we study the phylogenomic
287 evolution of a large commensal collection of *E. coli* (403 strains) over a 30-year period. This
288 collection is unique in its focus on commensals and the timespan, and we controlled as much as we
289 could for covariates.

290 The proportion of phylogroup B2 strains increased from 9% to 36% while the proportion of
291 phylogroup A strains decreased from 58% to 26% between 1980 and 2010, as already observed in
292 a large part of this collection (14). A predominance of B2 strains over other phylogroups in
293 commensal samples has also been reported in different industrialized countries. For instance, in the
294 late 1990s, the frequency of B2 strains in Australia, Japan, Sweden and USA is between 44 and
295 48% (1). It is possible that phylogroup B2 increased in frequency just before the 1990s in these
296 industrialized countries as they did in France from 1980 to 2010.

297 The diversity of STs and serotypes increased over time. The increase in ST diversity was
298 not explained by the increase in frequency of B2 strains (figure S3), but rather by the increase in
299 frequency of rare STs from 2001 to 2010.

300 We studied in more detail three major STs: two responsible for extra-intestinal infections
301 in humans (ST69 and ST95) (2, 37) and one (ST10) found at high frequency in the human gut as
302 well as in animals (38) (figure 2). The homoplasy (gain or loss of a character independently in
303 different lineages of the tree) of several surface antigens, especially in the ST10 (39), is in
304 agreement with their encoding genes being located in the major hot-spots of recombination in the
305 genome (4). Diversification per time unit of genes encoding for O and H antigens and of for FimH
306 protein for ST69 is faster than for ST10 and ST95, which is potentially explained by its younger
307 age as it can be expected during adaptive radiations for example, where lineages diversify rapidly
308 at the emergence of the clade (40).

309 Several O-groups have been considered as a promising target for a bioconjugate vaccine
310 against extra-intestinal infections (41). Recently, a phase 2 randomized controlled trial showed that
311 a vaccine targeting the O1, O2, O6, and O25-antigens was well tolerated and elicited an antibody
312 response against these antigens (8, 9). They represent a large part (24%) of our commensal strains,
313 suggesting the vaccine could perturb the commensal gut microbiota by eliminating commensal *E.*
314 *coli* clones in addition to the major pathogenic clones.

315

316 ***2. Recent temporal evolution of resistance and virulence in commensal* E. coli**

317 Gene and phylogroup frequencies varied in time (figures 1) (14, 34). Strains of phylogroup
318 B2, which carry many VFs, as well as individual VFs increased in frequency, whereas the observed
319 increase in resistance between 1980 and 2001 seems decoupled from phylogroups. To investigate
320 whether virulence and antibiotic resistance evolution are governed by different evolutionary

321 mechanisms, we decomposed the change in frequency of a gene into a term quantifying the clonal
322 expansion of STs carrying this gene and a term quantifying the increase in gene frequency within
323 STs.

324 The frequency of STs can vary in time, randomly or in response to selective pressures. As
325 a consequence, the frequency of a focal gene varies accordingly to the variation in frequency of the
326 STs carrying it (whether or not the gene is under selection). The change in gene frequency may
327 result from vertical transfers (clonal expansion) and horizontal transfers within STs. We called this
328 process ST driven change. The frequency of a gene can also vary concomitantly in one or several
329 STs regardless of variation in ST frequencies. This can result from horizontal transfers between or
330 within STs or the increase in frequency of the lineage carrying the focal gene at the expense of
331 others within STs, suggesting direct selection on the gene. We called this process gene driven
332 change. The contributions of gene driven and ST driven change varied between virulence and
333 resistance genes, and between the two periods (1980-2000 and 2000-2010) (figure 5).

334 Between 1980 and 2000, the virulence of *E. coli* strains rapidly evolved through both the
335 rise of more virulent STs and within-ST increase in frequency of virulence genes (gene driven
336 change) (figures 4A and 5A). In the second time-period from 2001 to 2010, the increase in
337 frequency of STs was mainly driven by the increase in frequency of more virulent STs (ST driven
338 change) (figures 4B and 5A). Significant changes in frequency suggest a role for direct selection
339 on virulence genes in driving the increased virulence. The presence of several of these genes may
340 be under positive directional selection. Moreover, the sequence diversity of many of them (37 out
341 of 55) also suggests selection, either diversifying or purifying.

342 What factors could explain the recent increase in virulence? Recent environmental changes
343 include a shift towards more processed, and more nutrient-dense food over the period 1969-2002
344 in France, and a stabilization over 2002-2010 (42), mirroring the spread of virulence genes within

345 ST observed in the period 1980-2000. This could select for virulence genes (e.g., iron acquisition
346 factors) or for virulent STs. Whether bacteria carrying some virulence genes are better adapted to
347 nutrient-dense diets could be experimentally tested (43). In an effort to do so, some of us recently
348 analyzed how mice diet affected the density of *E. coli* in the mice gut and found that the B2 strains
349 used was more prevalent in a high sugar high fat diet than in a high fiber diet (44). In addition to
350 nutriment availability, protection again host immune defense, phages and protozoans may also
351 drive the evolution of virulence (2, 45). A change in the diversity and abundance of these predators
352 might likewise explain the recent increase in virulence observed in commensal *E. coli*.

353 For resistance genes, the rapid increase in frequency of several genes [notably *aph*(6)-Id,
354 $bla_{\text{TEM-1B}}$ and *tet*(A), which are located on plasmid or transposons] (46, 47) from 1980 to 2000 was
355 mainly driven by an increase in the frequency of these genes independently within several STs
356 (figure 6A). Their presence is under positive selection and the gene sequence of $bla_{\text{TEM-1B}}$ and
357 *tet*(A) is under purifying selection. In contrast to 1980-2000, after 2001 the frequency of resistance
358 almost did not change (figures 5B and 6). Concomitantly, the proportion of strains resistant to two
359 or more antibiotic categories decreased, indicating a decrease in multi-resistance. In the early
360 2000s, levels of antibiotic resistance were higher in France than in most European countries (22,
361 25). A nationwide awareness campaign was launched in 2001 leading to a 26.5% reduction in
362 antibiotic use in humans over 5 years (25). The reduction in antibiotic use could explain the
363 stabilization in resistance gene frequencies observed in our data.

364 Our work has some limitations. First, only one isolate per subject was sampled. It is well
365 known that populations of subdominant clones are present in the feces (36). Resistant clones are
366 often subdominant and isolated using antibiotic containing plates. They have a specific population
367 structure and epidemiology (48). Second, our data set ends in 2010 and we did not capture the
368 recent evolution. A stool collect campaign is underway and will allow to address this point. Third,

369 the five collections gathered for this study are heterogeneous in terms of host age, sex, location
370 (Paris or Brittany) and potential (but unlikely) antibiotic use. However, we controlled for these
371 factors in our analyses, and we did not find any large effect of host age, location, sex and the
372 possibility of antibiotic consumption on gene content or genotype.
373
374 ***Conclusion***
375 To investigate the recent evolutionary dynamics of commensal *E. coli*, we whole-genome
376 sequenced a large collection of 403 strains sampled from the gut of healthy volunteers. Several
377 antibiotic resistance genes rapidly spread from 1980 to 2000, largely unhindered by clonal
378 structure. Higher virulence evolved through increase in virulence gene frequency within STs and
379 clonal expansion of more virulent STs. Increasing virulence of *E. coli* would result, everything else
380 being equal, in an increasing incidence of extra-intestinal infections, and could indeed contribute
381 to the increasing incidence of bacteremia observed in the last decades (12, 15). Future research
382 should investigate whether the observed increasing virulence in commensal strains was observed
383 in other geographical areas and what environmental factors could select for *E. coli* virulence.
384
385 **METHODS**
386 ***1. Strain collections***
387 We studied the whole genomes of four hundred and thirty-six *E. coli* strains gathered from stools
388 of 403 healthy adults (at least 18-year-old) of both sexes living in the Paris area or Brittany (both
389 locations in the North of France) between 1980 to 2010 (see main characteristics of the subjects in
390 table S3). These strains come from five previously published collections: VDG sampled in 1980
391 (49), ROAR in 2000 (50), LBC in 2001 (51), PAR in 2002 (51), and Coliville in 2010 (table S1)
392 (14). In all studies, one single *E. coli* colony randomly picked on the Drigalski plate was retained

393 per individual (14), probably representing the dominant strain. The study was approved by the
394 ethics evaluation committee of Institut National de la Santé et de la Recherche Médicale (INSERM)
395 (CCTIRS no. 09.243, CNIL no. 909277, and CQI no. 01-014).

396 In order to improve the temporal span for dating, we used 24 sequences from the Murray
397 collection with samples ranging from 1930 to 1941 (table S19) (52). From the 50 genomic
398 sequences available in the Murray collection, we selected sequences with available sampling time
399 and excluded multiple variants per strain (when strain name, date on tube and origin were identical
400 for two samples).

401

402 ***2. Sequencing of the commensal strains***

403 After DNA extraction, whole-genome of each strain was sequenced using Illumina NextSeq 2x150
404 bp after NextEra XT library preparation (Illumina, San Diego, CA) (26).

405

406 ***3. Assembly and typing***

407 The assembly of genomes was performed using the in-house script Petanc that integrates several
408 existing bacterial genomic tools (53), including SPADES (54). This in-house script was also used
409 to perform the typing of strains with several genotyping schemes using the genomic tool SRST2
410 (55). Multilocus sequence typing (MLST) was performed and STs were defined using the Warwick
411 MLST scheme and the Pasteur scheme (33, 56). We only used the Warwick scheme for the analyses
412 described thereafter. We also determined the O:H serotypes and the *fimH* alleles (7, 57). The
413 phylogroups were defined using the ClermonTyping method (58). The in-house script Petanc was
414 also used to type the Murray strains in order to determine which strains could be used for dating.

415

416 ***4. Distribution of phylogroups, STs and O-groups through time***

417 Because there were three time points close in time with a small number of strains, we aggregated
418 data of years 2000, 2001 and 2002 and it will henceforth be referred to as 2001.

419 To evaluate the frequency of phylogroups through time, we fitted a binomial generalized linear
420 model (GLM) for the presence/absence of each phylogroup as a function of time (continuous
421 variable), and controlling for several factors heterogeneous over the five collections: host age,
422 location (two categories: Paris vs. Brittany), sex and antibiotic consumption (two categories: "no"
423 vs. "unlikely"). When only an interval was available for age (113 strains from the collections PAR,
424 ROAR and VDG), we used the mean value of this interval. The same analyses were also performed
425 for STs and O-groups.

426 We generated stacked area charts of phylogroups, STs and O-groups to display their diversity
427 variation between 1980 and 2010.

428 We analyzed 53 strains in 1980, 104 between 2000 and 2002, and 246 in 2010. To determine
429 whether the number of distinct STs detected each year varied through time or was simply reflecting
430 the number of sampled strains, we generated null distributions of ST number depending on the
431 sampling effort. We sampled 10,000 times a fixed number of strains (53 for 1980, 104 for 2001
432 and 246 for 2010) with the frequency of each ST set to its overall frequency. For each time point,
433 the observed value was compared to the corresponding null distribution. The same analysis was
434 also performed for serotypes (O and H combinations), O-groups, H-types and *fimH* alleles.

435 To evaluate the influence of the change in frequency of phylogroup on ST diversity, we
436 generated null distributions of ST number depending on the phylogroup frequency and on the
437 sampling effort. We sampled 10,000 times a fixed number of strains corresponding to the number
438 of strains sampled by phylogroup and by year with the frequency of each ST set to its overall
439 frequency for each phylogroup. For each time point, the observed value was compared to the
440 corresponding null distribution.

441

442 ***5. Genomic diversity of the core genome***

443 The 403 assemblies (our commensal collection) were annotated with Prokka (59). We then
444 performed pan-genome analysis from annotated assemblies with Roary on the 403 genomes using
445 default parameters (60). The alignment of the core genome and the list of genes of the accessory
446 genome were generated.

447

448 ***6. Core genome phylogeny of* E. coli**

449 To build the phylogeny of *E. coli*, we aligned whole genomes (403 strains) to the reference
450 R1B5J10 (ST206 strain, close to the most frequent ST in our dataset, ST10) with Snippy 4.4.0
451 using standard parameters (61). We did not remove recombination events because it could
452 accentuate errors in phylogenetic distances and the topology of the tree is usually not affected (62).
453 The alignment of 266,665 SNPs (single-nucleotide polymorphism), was used to produce a
454 maximum likelihood phylogenetic tree with RAxML (63), using the GTRGAMMA model with
455 1,000 bootstrap replicates.

456

457 ***7. Divergence times estimates for major STs***

458 To further study the five most prevalent STs, ST10, ST73, ST95, ST69 and ST59, we aligned
459 whole genomes (136 strains from our collection and 10 from the Murray collection) to references,
460 R1B5J10, 016-002, R1B6J15, H1-004-0023-R-J and IAI39 respectively with Snippy 4.4.0 using
461 standard parameters (61). For each ST, three outgroup sequences were selected as the three closest
462 sequences to the focal ST belonging to two distinct and well supported clades, from the core
463 genome phylogenetic tree. Recombination events were excluded with Gubbins using default
464 settings (64). The five resulting alignments were used to produce a maximum likelihood

465 phylogenetic tree with RAxML using the GTRGAMMA model with 1,000 bootstrap replicates
466 (63).

467 The temporal analysis was performed with the program BEAST v1.10.4 (65), on three of the
468 most frequent STs with a molecular clock signal (i.e. positive relationship between root-to-tip
469 distance and time): ST10, ST95 and ST69. BEAST implements the "uncorrelated relaxed clock"
470 to model uncorrelated rate variations among lineages; the evolutionary rate of each branch is drawn
471 independently from a common underlying distribution. To estimate divergence times with this
472 model we used the following parameters: uncorrelated log-normal clock, substitution model
473 GTR+I+G, fixed topology (RAxML tree) and 500 million generations with sampling every 1000
474 generations and a burn-in of 50 million generations. We ran three replicates with three tree priors:
475 coalescent with constant population, coalescent with exponential growth and coalescent with
476 Bayesian skyline. The sample times were used to calibrate the tips. For each dataset, the best fitting
477 model set was determined by computing Bayes factors from marginal likelihoods estimations
478 calculated by stepping-stone sampling (66). Only the models that converged well and had and
479 effective sample size (ESS) larger than 200 for each parameter were compared. The best fitting
480 model was used in the subsequent analyses.

481

482 ***8. Virulence and resistance gene repertoire analyses***

483 The resistome, the virulome and the plasmid type were defined using the in-house script Petanc
484 (53). They were defined by BlastN with Abricate (https:// github.com/tseemann/abricate) using the
485 ResFinder database (67), a custom database including the VirulenceFinder database and VFDB
486 (68, 69), to which we added selected genes (table S20), and PlasmidFinder respectively (70). We
487 set the threshold for minimum identity to 80% and for minimum coverage to 90%. If several copies
488 of a gene were detected for a strain, we only kept the copy with the maximal sum of coverage and

489 identity. We also searched for point mutations responsible for betalactam (*ampC* promoter) and
490 fluroquinolone (*gyrA*/*B*, *parC*/*E*) resistance (67). The plasmid sequences were also predicted by
491 PlaScope (71).

492 To visually explore the relationships between the phylogroups and the resistance and
493 virulence gene repertory we performed a correspondence analysis (CA) (72) with the R package
494 'FactoMineR' (73). A CA is a multivariate graphical analysis used to explore the associations
495 among categorical variables. For each item in a table, a set of factor scores (coordinates) is obtained
496 from linear combinations of rows and columns. These coordinates are projected on two dimensions,
497 with in our case, the first axis opposing the phylogroups with the largest differences. The further
498 the gene categories are from the origin the more they are discriminating. If two variables are close
499 to each other in the plan, they are considered to be strongly associated. The significance of
500 associations between phylogroups and resistance and virulence categories was tested with a chi-
501 squared test.

502 For each isolate, we computed a virulence score corresponding to the number of the
503 virulence factors (VF) present in each isolate (number of products out of the 104 tested, table S20)
504 (adapted from (74)). The frequency of antibiotic resistance was computed for each year. Both gene
505 acquisitions and point mutations were included (table S21). The corresponding antibiotics were
506 retrieved from ResFinder (75). We next tested whether the virulence score and the frequency of
507 antibiotic resistance changed in time. We fitted a linear model for the number of VF products or
508 for the frequency of antibiotic resistance as a function of year (1980, 2001 (2000-2002), 2010) as
509 a categorical variable.

510 To decipher whether changes in gene frequency through time were due to variations in ST
511 frequencies or to variations in gene frequencies within STs, we decomposed the overall change in
512 gene frequency (Δf) as follows (figure S9). We call $f_{i,t}$ the frequency of the focal gene in ST i at

513 time *t*. We call $p_{i,t}$ the frequency of the ST *i* at time *t*. We are interested in the change in frequency

514 of the focal gene from time t_1 to time t_2. The change in gene frequency can be decomposed as

515 such:

516

517 $$\Delta f = \underbrace{\sum_{i=1}^{S} \Delta f_i \ \bar{p}_i}_{within\ ST} + \underbrace{\sum_{i=1}^{S} \Delta p_i \ \bar{f}_i}_{between\ ST},$$

518

519 where S is the number of STs, $\Delta f_i = f_{i,t_2} - f_{i,t_1}$ is the change in the focal gene frequency in ST *i*

520 from time t_1 to time t_2, $\Delta p_i = p_{i,t_2} - p_{i,t_1}$ the change in the frequency of ST *i* from time t_1 to time

521 t_2, $\bar{p}_i = \frac{1}{2}\left(p_{i,t_1} + p_{i,t_2}\right)$ the mean ST *i* frequency in the sample (weighting the two timepoints

522 equally) and $\bar{f}_i = \frac{1}{2}\left(f_{i,t_1} + f_{i,t_2}\right)$ the mean gene frequency in ST *i* in the sample.

523 For virulence, when the frequency of several alleles for a single gene was assessed by our

524 in-house script, we added these frequencies for each corresponding gene (table S20).

525 The decomposed temporal changes of virulence and of antibiotic resistance were

526 summarized as follow. For virulence, we first computed the mean change per product, and then the

527 mean change among product for each category (see table S20). For resistance, we computed the

528 mean change among genes for each category (see table S21).

529 For each time period, and for each gene, we used Fisher's exact test to examine the

530 significance of the association between gene presence/absence and two consecutive timepoints

531 (1980-2001 or 2001-2010).

532 In addition, we tested if the genes with a significant change in frequency over at least one

533 of the two periods (according to Fisher's exact test) were under directional selection over the whole

534 period 1980-2010. More precisely we tested 1) if the presence of these genes (relative to absence)

535 was under selection and 2) if there was a sign of selection acting on the sequence of these genes.
536 First, we fitted a logistic model for gene presence/absence as a function of time (continuous
537 variable, from 1980 to 2020) and phylogroup (categorial variable). Second, we computed dN/dS
538 ratios. We first aligned sequences of each gene with mafft (76). We generated the consensus
539 sequence of each gene (function 'ConsensusSequence' in package DECIPHER in R) (77). Then,
540 we calculated dN/dS ratios between each unique sequence and the consensus sequence (function
541 'dnds' in package ape in R) (78). Finally, we tested if the mean dN/dS ratio obtained for each gene
542 was significantly different from 1 (t-test). We could not compute the dN/dS ratio for six genes
543 because there were less than 3 different sequences in the alignment (*kps*T, *stx*1A, *stx*1B, *sul*1) or
544 because there was no synonymous mutation (*aph*(6)-Id and *hly*F).
545
546 **DECLARATIONS**
547 **Ethics approval and consent to participate**
548 Not applicable
549
550 **Data availability**
551 The data generated in this study have been submitted to the NCBI BioProject database under the
552 accession numbers PRJEB39252 (Coliville)
553 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB39252], PRJEB38489 (ROAR)
554 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB38489], PRJEB44819 (LBC)
555 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB44819], PRJEB44872 (PAR)
556 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB44872], PRJEB44873 (VDG)
557 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB44873].
558

Disclosure of interest

The authors report no conflict of interest.

Acknowledgements

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing resources. JM and FB were funded by the CNRS Momentum grant to FB. ED was partially supported by the "Fondation pour la Recherche Médicale" (Equipe FRM 2016, grant number DEQ20161136698). GR was supported by a "Poste d'accueil" funded by the Assistance-Publique Hôpitaux de Paris (AP-HP) and the Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) personal grant for his PhD.

Appendixes

Tables S1-21 (SupplementalTables_S1_21.xlsx)

Figures S1-9 (SupplementalFigures_S1_9.docx)

575 **REFERENCES**

576 1. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal
577 *Escherichia coli*. 3. Nat Rev Microbiol 8:207–217.
578 2. Denamur E, Clermont O, Bonacorsi S, Gordon D. 2020. The population genetics of
579 pathogenic *Escherichia coli*. 1. Nat Rev Microbiol 19:37–54.
580 3. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, Han C, Bisignano C,
581 Rao P, Wool E, Johnson SC, Browne AJ, Chipeta MG, Fell F, Hackett S, Haines-Woodhouse
582 G, Hamadani BHK, Kumaran EAP, McManigal B, Agarwal R, Akech S, Albertson S, Amuasi
583 J, Andrews J, Aravkin A, Ashley E, Bailey F, Baker S, Basnyat B, Bekker A, Bender R,
584 Bethou A, Bielicki J, Boonkasidecha S, Bukosia J, Carvalheiro C, Castañeda-Orjuela C,
585 Chansamouth V, Chaurasia S, Chiurchiù S, Chowdhury F, Cook AJ, Cooper B, Cressey TR,
586 Criollo-Mora E, Cunningham M, Darboe S, Day NPJ, Luca MD, Dokova K, Dramowski A,
587 Dunachie SJ, Eckmanns T, Eibach D, Emami A, Feasey N, Fisher-Pearson N, Forrest K,
588 Garrett D, Gastmeier P, Giref AZ, Greer RC, Gupta V, Haller S, Haselbeck A, Hay SI, Holm
589 M, Hopkins S, Iregbu KC, Jacobs J, Jarovsky D, Javanmardi F, Khorana M, Kissoon N,
590 Kobeissi E, Kostyanev T, Krapp F, Krumkamp R, Kumar A, Kyu HH, Lim C,
591 Limmathurotsakul D, Loftus MJ, Lunn M, Ma J, Mturi N, Munera-Huertas T, Musicha P,
592 Mussi-Pinhata MM, Nakamura T, Nanavati R, Nangia S, Newton P, Ngoun C, Novotney A,
593 Nwakanma D, Obiero CW, Olivas-Martinez A, Olliaro P, Ooko E, Ortiz-Brizuela E, Peleg
594 AY, Perrone C, Plakkal N, Ponce-de-Leon A, Raad M, Ramdin T, Riddell A, Roberts T,
595 Robotham JV, Roca A, Rudd KE, Russell N, Schnall J, Scott JAG, Shivamallappa M,
596 Sifuentes-Osornio J, Steenkeste N, Stewardson AJ, Stoeva T, Tasak N, Thaiprakong A,
597 Thwaites G, Turner C, Turner P, Doorn HR van, Velaphi S, Vongpradith A, Vu H, Walsh T,
598 Waner S, Wangrangsimakul T, Wozniak T, Zheng P, Sartorius B, Lopez AD, Stergachis A,

599 Moore C, Dolecek C, Naghavi M. 2022. Global burden of bacterial antimicrobial resistance in
600 2019: a systematic analysis. The Lancet 399:629–655.
601 4. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S,
602 Bouchier C, Bouvet O. 2009. Organised genome dynamics in the *Escherichia coli* species
603 results in highly diverse adaptive paths. PLoS Genet 5.
604 5. Orskov I, Orskov F, Jann B, Jann K. 1977. Serology, chemistry, and genetics of O and K
605 antigens of *Escherichia coli*. Bacteriol Rev 41:667–710.
606 6. Schembri MA, Kjaergaard K, Sokurenko EV, Klemm P. 2001. Molecular characterization of
607 the *Escherichia coli* FimH adhesin. J Infect Dis 183:S28–S31.
608 7. Roer L, Tchesnokova V, Allesøe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen
609 MCF, Lund O, Hansen F, Hammerum AM, Sokurenko E, Hasman H. 2017. Development of a
610 web tool for *Escherichia coli* subtyping based on *fimH* alleles. J Clin Microbiol 55:2538–
611 2543.
612 8. Frenck RW, Ervin J, Chu L, Abbanat D, Spiessens B, Go O, Haazen W, van den Dobbelsteen
613 G, Poolman J, Thoelen S, Ibarra de Palacios P. 2019. Safety and immunogenicity of a vaccine
614 for extra-intestinal pathogenic *Escherichia coli* (ESTELLA): a phase 2 randomised controlled
615 trial. Lancet Infect Dis 19:631–640.
616 9. Huttner A, Hatz C, van den Dobbelsteen G, Abbanat D, Hornacek A, Frölich R, Dreyer AM,
617 Martin P, Davies T, Fae K, van den Nieuwenhof I, Thoelen S, de Vallière S, Kuhn A,
618 Bernasconi E, Viereck V, Kavvadias T, Kling K, Ryu G, Hülder T, Gröger S, Scheiner D,
619 Alaimo C, Harbarth S, Poolman J, Fonck VG. 2017. Safety, immunogenicity, and preliminary
620 clinical efficacy of a vaccine against extraintestinal pathogenic *Escherichia coli* in women
621 with a history of recurrent urinary tract infection: a randomised, single-blind, placebo-
622 controlled phase 1b trial. Lancet Infect Dis 17:528–537.

623 10. Johnson JR. 1991. Virulence factors in *Escherichia coli* urinary tract infection. Clin Microbiol
624 Rev 4:80–128.
625 11. Clermont O, Couffignal C, Blanco J, Mentré F, Picard B, Denamur E, Groups the C and C.
626 2017. Two levels of specialization in bacteraemic *Escherichia coli* strains revealed by their
627 comparison with commensal strains. Epidemiol Infect 145:872–882.
628 12. Kraker MEA de, Davey PG, Grundmann H, Group on behalf of the B study. 2011. Mortality
629 and hospital stay associated with resistant *Staphylococcus aureus* and *Escherichia coli*
630 bacteremia: estimating the burden of antibiotic resistance in Europe. PLOS Med 8:e1001104.
631 13. Didelot X, Darling A, Falush D. 2009. Inferring genomic flux in bacteria. Genome Res
632 19:306–317.
633 14. Massot M, Daubié A-S, Clermont O, Jaureguy F, Couffignal C, Dahbi G, Mora A, Blanco J,
634 Branger C, Mentré F, Eddi A, Picard B, Denamur E. 2016. Phylogenetic, virulence and
635 antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from
636 community subjects in the Paris area in 2010 and evolution over 30 years. Microbiology
637 162:642–650.
638 15. Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ, Dunn L, Jeffery K,
639 Butler CC, Hayward G, Andersson M, Morgan M, Oakley S, Mason A, Hopkins S, Wyllie
640 DH, Crook DW, Wilcox MH, Johnson AP, Peto TEA, Walker AS. 2018. Trends over time in
641 *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities
642 in Oxfordshire, UK, 1998–2016: a study of electronic health records. Lancet Infect Dis
643 18:1138–1149.
644 16. Birgy A, Levy C, Bidet P, Thollot F, Derkx V, Béchet S, Mariani-Kurkdjian P, Cohen R,
645 Bonacorsi S. 2016. ESBL-producing Escherichia coli ST131 versus non-ST131: evolution and

646 risk factors of carriage among French children in the community between 2010 and 2015. J
647 Antimicrob Chemother 71:2949–2956.
648 17. Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O. 2007.
649 Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic
650 group *Escherichia coli* strains. Mol Biol Evol 24:2373–2384.
651 18. Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. 2010. Pathogenicity-associated
652 islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal
653 colonization. J Bacteriol 192:4885–4893.
654 19. Östblom A, Adlerberth I, Wold AE, Nowrouzian FL. 2011. Pathogenicity island markers,
655 virulence determinants malX and usp, and the capacity of *Escherichia coli* to persist in
656 infants' commensal microbiotas. Appl Environ Microbiol 77:2303–2308.
657 20. Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. 2014. *Escherichia coli* ST131, an intriguing
658 clonal group. Clin Microbiol Rev 27:543–574.
659 21. Martin P, Marcq I, Magistro G, Penary M, Garcie C, Payros D, Boury M, Olier M,
660 Nougayrède J-P, Audebert M, Chalut C, Schubert S, Oswald E. 2013. Interplay between
661 siderophores and colibactin genotoxin biosynthetic pathways in *Escherichia coli*. PLOS
662 Pathog 9:e1003437.
663 22. Goossens H, Ferech M, Vander Stichele R, Elseviers M. 2005. Outpatient antibiotic use in
664 Europe and association with resistance: a cross-national database study. The Lancet 365:579–
665 587.
666 23. Chatterjee A, Modarai M, Naylor NR, Boyd SE, Atun R, Barlow J, Holmes AH, Johnson A,
667 Robotham JV. 2018. Quantifying drivers of antibiotic resistance in humans: a systematic
668 review. Lancet Infect Dis 18:e368–e378.

669 24. Tedijanto C, Olesen SW, Grad YH, Lipsitch M. 2018. Estimating the proportion of bystander
670 selection for antibiotic resistance among potentially pathogenic bacterial flora. Proc Natl Acad
671 Sci 115:E11988–E11995.
672 25. Sabuncu E, David J, Bernède-Bauduin C, Pépin S, Leroy M, Boëlle P-Y, Watier L, Guillemot
673 D. 2009. Significant reduction of antibiotic use in the community after a nationwide campaign
674 in France, 2002–2007. PLoS Med 6:e1000084.
675 26. de Lastours V, Laouénan C, Royer G, Carbonnelle E, Lepeule R, Esposito-Farèse M,
676 Clermont O, Duval X, Fantin B, Mentré F, Decousser JW, Denamur E, Lefort A, Group on
677 behalf of the S. 2020. Mortality in *Escherichia coli* bloodstream infections: antibiotic
678 resistance still does not make it. J Antimicrob Chemother 75:2334–2343.
679 27. Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill
680 J. 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates
681 a stable population structure only transiently disturbed by the emergence of ST131. Genome
682 Res 27:1437–1449.
683 28. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C,
684 Cookson BT, Shendure J. 2015. Large-scale genomic sequencing of extraintestinal pathogenic
685 *Escherichia coli* strains. Genome Res 25:119–128.
686 29. Foxman B. 2010. The epidemiology of urinary tract infection. 12. Nat Rev Urol 7:653–660.
687 30. Goto M, McDanel JS, Jones MM, Livorsi DJ, Ohl ME, Beck BF, Richardson KK, Alexander
688 B, Perencevich EN. 2017. Antimicrobial nonsusceptibility of gram-negative bloodstream
689 isolates, Veterans Health Administration System, United States, 2003–20131. Emerg Infect
690 Dis 23:1815–1825.
691 31. Kauffmann F. 1947. The serology of the *coli* group. J Immunol Baltim Md 1950 57:71–100.

692 32. Arimizu Y, Kirino Y, Sato MP, Uno K, Sato T, Gotoh Y, Auvray F, Brugere H, Oswald E,
693 Mainil JG, Anklam KS, Döpfer D, Yoshino S, Ooka T, Tanizawa Y, Nakamura Y, Iguchi A,
694 Morita-Ishihara T, Ohnishi M, Akashi K, Hayashi T, Ogura Y. 2019. Large-scale genome
695 analysis of bovine commensal *Escherichia coli* reveals that bovine-adapted *E. coli* lineages are
696 serving as evolutionary sources of the emergence of human intestinal pathogenic strains.
697 Genome Res 29:1495–1505.
698 33. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ,
699 Ochman H, Achtman M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary
700 perspective. Mol Microbiol 60:1136–1151.
701 34. Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon
702 D, Rocha EP. 2020. Phylogenetic background and habitat drive the genetic diversification of
703 *Escherichia coli*. PLOS Genet 16:e1008866.
704 35. Qin X, Hu F, Wu S, Ye X, Zhu D, Zhang Y, Wang M. 2013. Comparison of adhesin genes
705 and antimicrobial susceptibilities between uropathogenic and intestinal commensal
706 *Escherichia coli* strains. PLOS ONE 8:e61169.
707 36. Smati M, Clermont O, Gal FL, Schichmanoff O, Jauréguy F, Eddi A, Denamur E, Picard B,
708 for the Coliville Group. 2013. Real-time PCR for quantitative analysis of human commensal
709 *Escherichia coli* populations reveals a high frequency of subdominant phylogroups. Appl
710 Environ Microbiol 79:5005–5012.
711 37. Basmaci R, Bonacorsi S, Bidet P, Biran V, Aujard Y, Bingen E, Béchet S, Cohen R, Levy C.
712 2015. *Escherichia coli* meningitis features in 325 children from 2001 to 2013 in France. Clin
713 Infect Dis 61:779–786.

714 38. Manges AR, Harel J, Masson L, Edens TJ, Portt A, Reid-Smith RJ, Zhanel GG, Kropinski
715 AM, Boerlin P. 2015. Multilocus sequence typing and virulence gene profiles associated with
716 *Escherichia coli* from human and animal sources. Foodborne Pathog Dis 12:302–310.
717 39. Royer G, Darty MM, Clermont O, Condamine B, Laouenan C, Decousser J-W, Vallenet D,
718 Lefort A, de Lastours V, Denamur E, Wolff M, Alavoine L, Duval X, Skurnik D, Woerther P-
719 L, Andremont A, Carbonnelle E, Lortholary O, Nassif X, Abgrall S, Jaureguy F, Picard B,
720 Houdouin V, Aujard Y, Bonacorsi S, Meybeck A, Barnaud G, Branger C, Lefort A, Fantin B,
721 Bellier C, Bert F, Nicolas-Chanoine M-H, Page B, Cremniter J, Gaillard J-L, Leturdu F, Sollet
722 J-P, Plantefève G, Panhard X, Mentré F, Marcault E, Tubach F, Zarrouk V, Bert F, Duprilot
723 M, Leflon-Guibout V, Maataoui N, Armand L, Nguyen LL, Collarino R, Munier A-L,
724 Jacquier H, Lecorché E, Coutte L, Gomart C, Fateh OA, Landraud L, Messika J, Aslangul E,
725 Gerin M, Bleibtreu A, Lescat M, Walewski V, Mechaï F, Dollat M, Maherault A-C, Wolff M,
726 Mercier-Darty M, Basse B, COLIBAFI and SEPTICOLI groups. 2021. Phylogroup stability
727 contrasts with high within sequence type complex dynamics of *Escherichia coli* bloodstream
728 infection isolates over a 12-year period. Genome Med 13:77.
729 40. Barrier M, Robichaux RH, Purugganan MD. 2001. Accelerated regulatory gene evolution in
730 an adaptive radiation. Proc Natl Acad Sci 98:10208–10213.
731 41. Poolman JT, Wacker M. 2016. Extraintestinal pathogenic *Escherichia coli*, a common human
732 pathogen: challenges for vaccine development and progress in the field. J Infect Dis 213:6–13.
733 42. Caillavet F, Darmon N, Létoile F, Nichèle V. 2018. Is nutritional quality of food-at-home
734 purchases improving? 1969–2010: 40 years of household consumption surveys in France. Eur
735 J Clin Nutr 72:220–227.
736 43. O'Brien CL, Gordon DMY 2011. 2011. Effect of diet and gut dynamics on the establishment
737 and persistence of *Escherichia coli*. Microbiology 157:1375–1384.

738 44. Ghalayini M, Magnan M, Dion S, Zatout O, Bourguignon L, Tenaillon O, Lescat M. 2019.
739 Long-term evolution of the natural isolate of *Escherichia coli* 536 in the mouse gut colonized
740 after maternal transmission reveals convergence in the constitutive expression of the lactose
741 operon. Mol Ecol 28:4470–4485.
742 45. Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG. 2004. Protozoan predation,
743 diversifying selection, and the evolution of antigenic diversity in *Salmonella*. Proc Natl Acad
744 Sci 101:10644–10649.
745 46. Khezri A, Avershina E, Ahmad R. 2021. Plasmid identification and plasmid-mediated
746 antimicrobial gene detection in Norwegian isolates. 1. Microorganisms 9:52.
747 47. Ribera A, Roca I, Ruiz J, Gibert I, Vila J. 2003. Partial characterization of a transposon
748 containing the tet(A) determinant in a clinical isolate of *Acinetobacter baumannii*. J
749 Antimicrob Chemother 52:477–480.
750 48. Day MJ, Hopkins KL, Wareham DW, Toleman MA, Elviss N, Randall L, Teale C, Cleary P,
751 Wiuff C, Doumith M, Ellington MJ, Woodford N, Livermore DM. 2019. Extended-spectrum
752 β-lactamase-producing *Escherichia coli* in human-derived and foodchain-derived samples
753 from England, Wales, and Scotland: an epidemiological surveillance and typing study. Lancet
754 Infect Dis 19:1325–1335.
755 49. Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventré A, Elion J, Picard B, Denamur E.
756 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among
757 geographically distinct human populations. Microbiology, 147:1671–1676.
758 50. Skurnik D, Clermont O, Guillard T, Launay A, Danilchanka O, Pons S, Diancourt L, Lebreton
759 F, Kadlec K, Roux D, Jiang D, Dion S, Aschard H, Denamur M, Cywes-Bentley C, Schwarz
760 S, Tenaillon O, Andremont A, Picard B, Mekalanos J, Brisse S, Denamur E. 2016. Emergence

761 of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. Mol Biol
762 Evol 33:898–914.
763 51. Escobar-Páramo P, Grenet K, Menac'h AL, Rode L, Salgado E, Amorin C, Gouriou S, Picard
764 B, Rahimy MC, Andremont A, Denamur E, Ruimy R. 2004. Large-scale population Structure
765 of human commensal *Escherichia coli* isolates. Appl Environ Microbiol 70:5698–5700.
766 52. Baker KS, Burnett E, McGregor H, Deheer-Graham A, Boinett C, Langridge GC, Wailan AM,
767 Cain AK, Thomson NR, Russell JE, Parkhill J. 2015. The Murray collection of pre-antibiotic
768 era Enterobacteriacae: a unique research resource. Genome Med 7:97.
769 53. Bourrel AS, Poirel L, Royer G, Darty M, Vuillemin X, Kieffer N, Clermont O, Denamur E,
770 Nordmann P, Decousser J-W, IAME Resistance Group. 2019. Colistin resistance in Parisian
771 inpatient faecal *Escherichia coli* as the result of two distinct evolutionary pathways. J
772 Antimicrob Chemother 74:1521–1530.
773 54. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM,
774 Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G,
775 Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its
776 applications to single-cell sequencing. J Comput Biol 19:455–477.
777 55. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014.
778 SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome
779 Med 6:90.
780 56. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E,
781 Lortholary O, Clermont O, Denamur E, Picard B, Nassif X, Brisse S. 2008. Phylogenetic and
782 genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics 9:560.
783 57. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, Levine MM, Robins-
784 Browne RM, Holt KE. 2016. In silico serotyping of E. coli from short read data identifies

785 limited novel O-loci but extensive diversity of O:H serotype combinations within and between
786 pathogenic lineages. Microb Genomics 2.
787 58. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. 2018. ClermonTyping:
788 an easy-to-use and accurate in silico method for *Escherichia genus* strain phylotyping. Microb
789 Genomics 4.
790 59. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–
791 2069.
792 60. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D,
793 Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis.
794 Bioinformatics 31:3691–3693.
795 61. Seemann T. 2015. Snippy: rapid haploid variant calling and core SNP phylogeny. GitHub
796 Available Github Comtseemannsnippy.
797 62. Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is
798 robust to recombination but demographic inference is not. mBio 5.
799 63. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
800 large phylogenies. Bioinformatics 30:1312–1313.
801 64. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR.
802 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome
803 sequences using Gubbins. Nucleic Acids Res 43:e15–e15.
804 65. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti
805 and the BEAST 1.7. Mol Biol Evol 29:1969–1973.
806 66. Baele G, Lemey P, Suchard MA. 2016. Genealogical working distributions for bayesian model
807 testing with phylogenetic uncertainty. Syst Biol 65:250–264.

808 67. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM,
809 Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob
810 Chemother 67:2640–2644.
811 68. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset
812 for big data analysis--10 years on. Nucleic Acids Res 44:D694-697.
813 69. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014.
814 Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection
815 of verotoxigenic *Escherichia coli*. J Clin Microbiol 52:1501–1510.
816 70. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller
817 Aarestrup F, Hasman H. 2014. *In Silico* detection and typing of plasmids using PlasmidFinder
818 and Plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903.
819 71. Royer G, Decousser JW, Branger C, Médigue C, Denamur E, Vallenet D. 2018. PlaScope: a
820 targeted approach to assess the plasmidome of *Escherichia coli* strains. preprint,
821 Bioinformatics.
822 72. Benzecri J. 1992. Correspondence analysis handbook. Marcel Decker, New York.
823 73. Lê S, Josse J, Husson F. 2008. FactoMineR : An *R* Package for multivariate analysis. J Stat
824 Softw 25.
825 74. Lefort A, Panhard X, Clermont O, Woerther P-L, Branger C, Mentré F, Fantin B, Wolff M,
826 Denamur E. 2011. Host factors and portal of entry outweigh bacterial determinants to predict
827 the severity of *Escherichia coli* bacteremia. J Clin Microbiol 49:777–783.
828 75. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL,
829 Rebelo AR, Florensa AF, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK,
830 Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum
831 MF, Duggett NA, Kempf I, Nykäsenoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L,

832 Mossong J, Losch S, Ragimbeau C, Lund O, Aarestrup FM. 2020. ResFinder 4.0 for
833 predictions of phenotypes from genotypes. J Antimicrob Chemother 75:3491–3500.
834 76. Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software Version 7:
835 Improvements in performance and usability. Mol Biol Evol 30:772–780.
836 77. Wright E S. 2016. Using DECIPHER v2.0 to Analyze big biological sequence data in R. R J
837 8:352.
838 78. Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and
839 evolutionary analyses in R. Bioinformatics 35:526–528.
840
841

842 **FIGURE LEGENDS**

843 **Figure 1.** Frequency distribution of phylogroups and STs between 1980 and 2010. (A) Frequency
844 distribution of phylogroups. The decline of phylogroup A and the increase of phylogroup B2 were
845 significant (0.05 level). (B) Frequency distribution of STs. The proportion of each ST has been
846 plotted by year ordered by the overall frequency (most common at the bottom). Only the names of
847 STs with an overall frequency > 0.3 and ST131 are shown. The only variation that was significant
848 at the 0.05 level was the decline of ST10. (C) Frequency of the O-groups O1, O2, O6 and O25. No
849 significant variation with time was detected (0.05 level).

850

851 **Figure 2.** Genomic content of three of the most frequent STs. For ST10 (A), ST69 (B) and ST95
852 (C), in addition to the epidemiological type diversity (O:H serotypes and *fimH* allele), we examined
853 the presence of five antibiotic resistance genes ($bla_{\text{TEM-1B}}$, *tet*(B), *tet*(A), *aph*(3'')-Ib and *aph*(6)-
854 Id) and of 34 virulence genes [*pap* (adhesin), *chu*, *iro*, *iuc* (iron capture systems) and *kps* (protectin)
855 operons]. The timetrees were built with BEAST v1.10.4 (65). ST10 exhibits the largest serotype
856 diversity (44 serotypes), among the other STs examined here, ST69 (4 serotypes) and ST95 (6
857 serotypes). Nodes with a support value (Bayesian posterior probability) > 0.75 are indicated by
858 black circles.

859

860 **Figure 3.** Evolution of virulence and antibiotic resistance between 1980 and 2010. (A) Mean
861 virulence score computed for all the 403 commensal *E. coli* isolates. The virulence score of a strain
862 is defined as the number of the VFs of each category tested that were present in that strain. (B)
863 Frequency of antibiotic resistance through time for all 403 isolates. Both gene acquisition and point
864 mutations are included, but we omitted the macrolide category because more than 99% of the

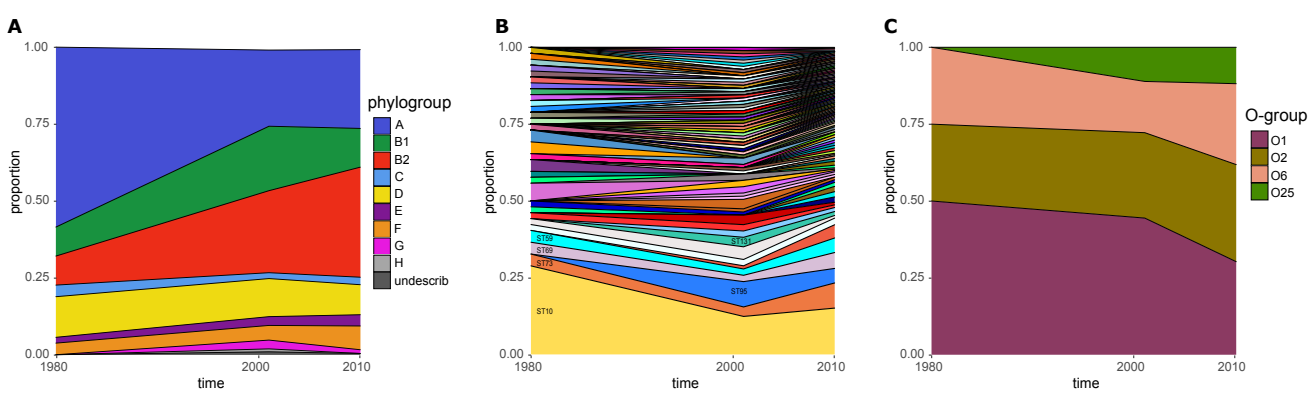
865 strains were resistant to macrolide antibiotics. For readability, in each panel the year 2001 includes
866 strains sampled in 2000, 2001 and 2002.
867
868 **Figure 4.** Temporal change of virulence frequency between 1980 – 2001 (A) and between 2001 –
869 2010 (B). The overall frequency change (Δf) for each gene (increases depicted by circles and
870 decreases by squares) is decomposed in change driven by the variation in frequency of STs carrying
871 the focal gene (ST driven change) and change driven by the variation in frequency of the focal gene
872 (gene driven change). For readability, only genes for which between ST change or within ST
873 change was greater than 0.02 in absolute value are shown (see table S13-14 for the complete list).
874 Genes highlighted in bold are those for which the temporal change is significant at the 0.05 level.
875 In the panel A, *chu** includes *chuA*, *chuS*, *chuT*, *chuU*, *chuV*, *chuW*, *chuX* and *chuY*.
876
877 **Figure 5.** Summary of the decomposed temporal change of virulence and antibiotic resistance
878 between 1980 and 2010. For virulence (A), we computed the mean change per product (e.g. mean
879 change among *iucA*, *iucB*, *iucC*, *iucD* and *iutA* genes for aerobactin), and reported the mean change
880 among products of a category (e.g. among products classified in the iron acquisition category). For
881 resistance (B), we computed the mean change among genes belonging to a given category (e.g.
882 among *tet*(A), *tet*(B), *tet*(D) and *tet*(M) for tetracycline).
883
884 **Figure 6.** Temporal change of antibiotic resistance frequency between 1980 – 2001 (A) and
885 between 2001 – 2010 (B). The overall frequency change (Δf) for each gene (increases depicted by
886 circles and decreases by squares) is decomposed in change driven by the variation in frequency of
887 STs carrying the focal gene (ST driven change) and change driven by the variation in frequency of
888 the focal gene (gene driven change). For readability, only genes for which between ST change or

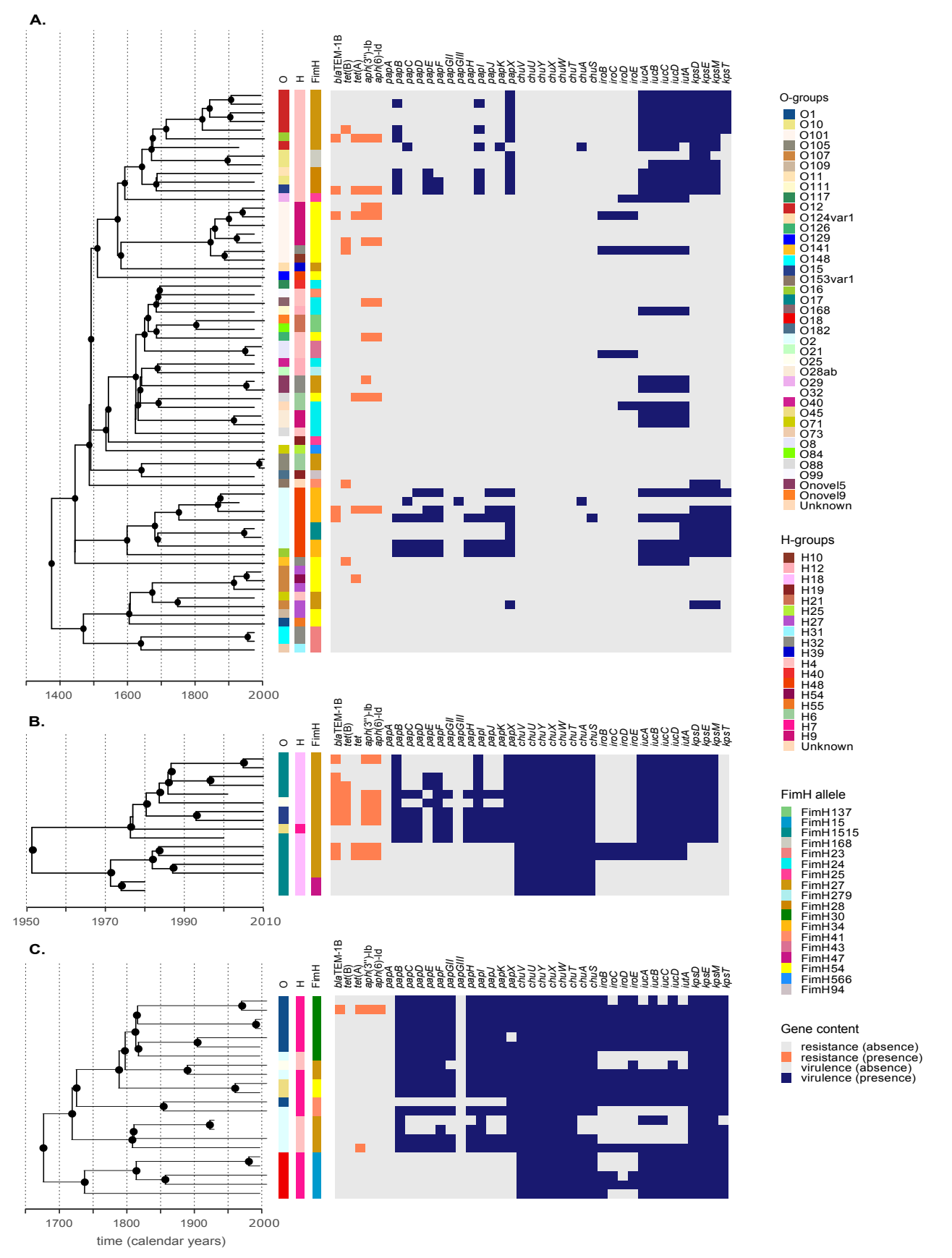
889 within ST change was greater than 0.02 are shown here (see table S15-16 for the complete list).
890 Genes highlighted in bold are those for which the temporal change is significant at the 0.05 level.
891 Note the scale of the y-axis is much smaller for resistance than for virulence (figure 3) as ST-driven
892 changes are minor compared to gene-driven changes for antibiotic resistance genes.
893

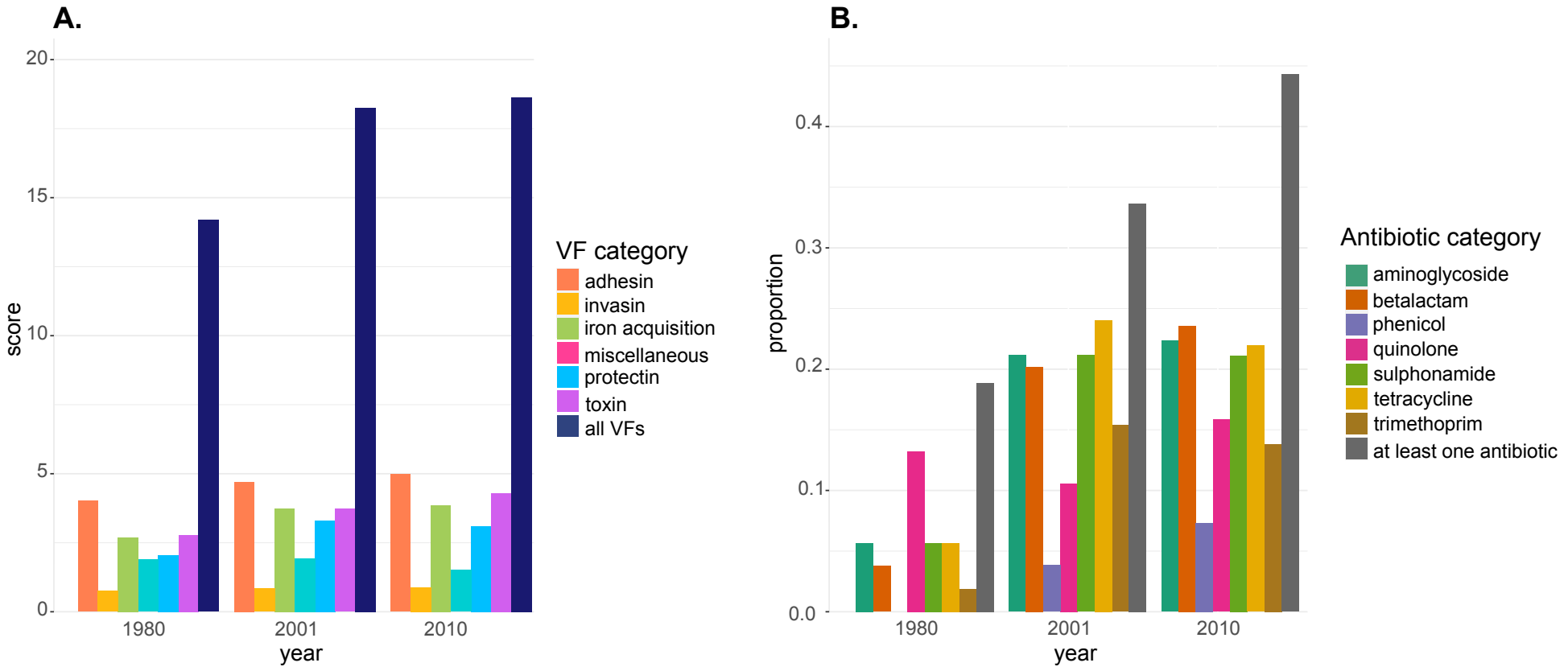
894 **Table 1**. **Distribution of the 11 most frequent sequence types of the *E. coli* commensal collection isolates.** The
895 number of isolates and the percentage for each year are presented in the table (see table S12 for the complete table).
896 The year 2001 corresponds to the aggregated data of years 2000, 2001 and 2002. The * indicate the sequence types
897 (STs) for which we inferred the divergence times.

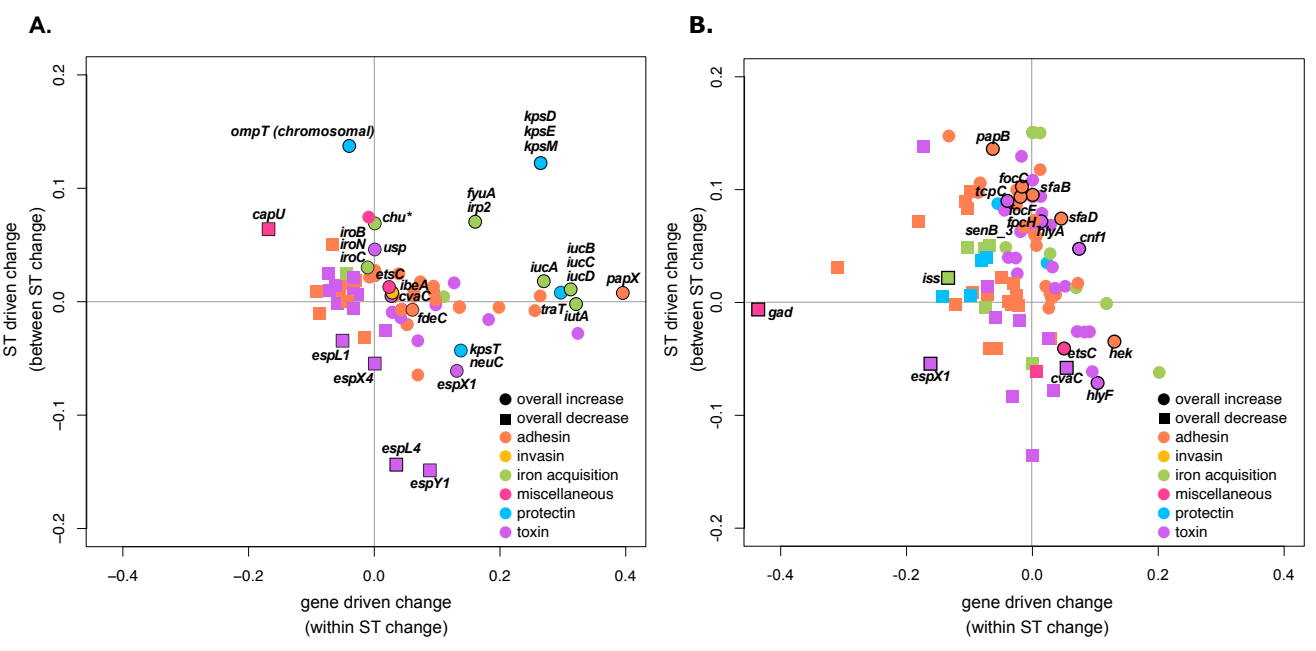
	Commensal			
ST (phylogroup)	**1980**	**2001**	**2010**	**All years**
10* (A)	15 (29%)	12 (12%)	35 (15%)	62 (16%)
73 (B2)	2 (4%)	3 (3%)	19 (8%)	24 (6%)
95* (B2)	0 (0%)	8 (8%)	11 (5%)	19 (5%)
69* (D)	2 (4%)	2 (2%)	12 (5%)	16 (4%)
59 (F)	2 (4%)	2 (2%)	11 (5%)	15 (4%)
141 (B2)	0 (0%)	1 (1%)	10 (4%)	11 (3%)
93 (A)	1 (2%)	4 (4%)	2 (1%)	7 (2%)
452 (B2)	1 (2%)	2 (2%)	5 (2%)	8 (2%)
405 (D)	1 (2%)	2 (2%)	2 (1%)	5 (1%)
58 (B1)	0 (0%)	2 (2%)	3 (1%)	5 (1%)
131 (B2)	0 (0%)	3 (3%)	3 (1%)	6 (2%)
Total in collections	53	104	246	403

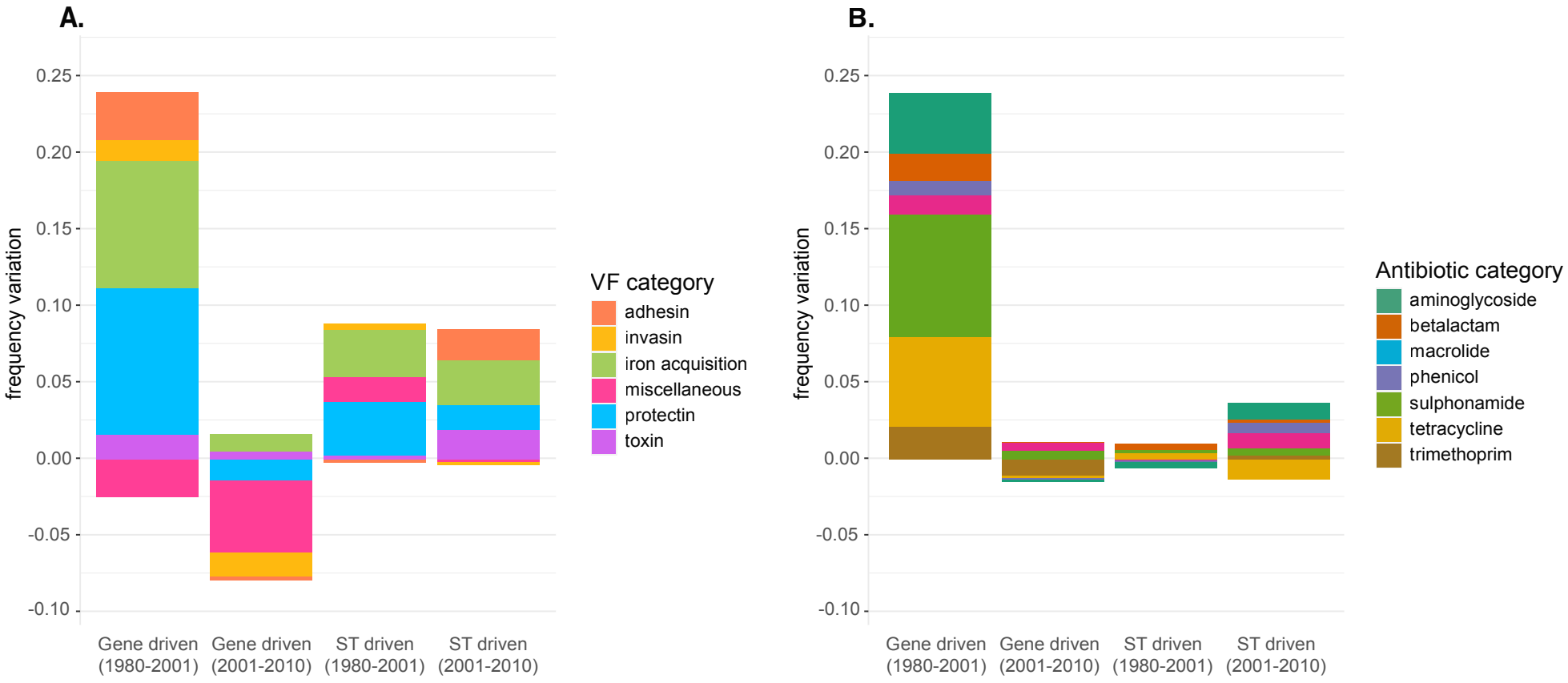
898

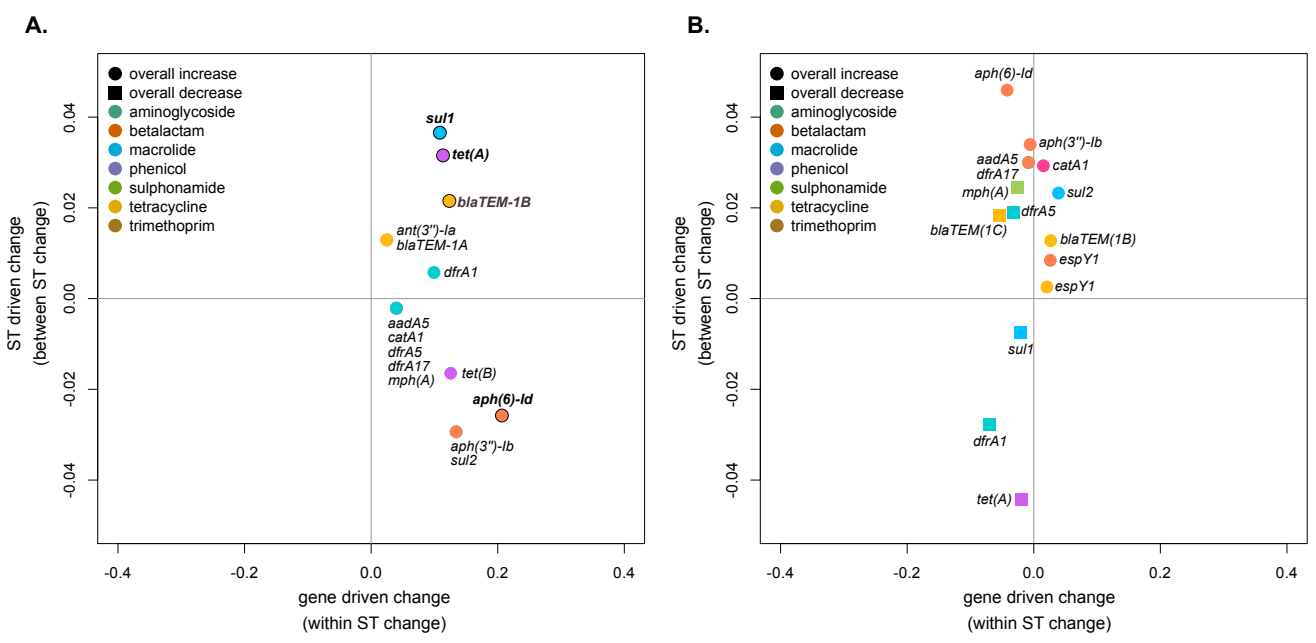

A
1.00
0.75
0.50
0.25
0.00
proportion
1980
2000
2010
time
phylogroup
A
B1
B2
C
D
E
F
G
H
undescrib
B
1.00
0.75
0.50
0.25
0.00
proportion
1980
2000
2010
time
ST69
ST73
ST131
ST95
ST10
C
1.00
0.75
0.50
0.25
0.00
proportion
1980
2000
2010
time
O-group
O1
O2
O6
O25


A.
B.
C.
O
H
FimH
blaTEM-1B
tet(B)
tet(A)
aph(3'')-Ib
aph(6)-Id
papA
papB
papC
papD
papE
papF
papGII
papGIII
papH
papI
papJ
papK
papX
chuV
chuU
chuY
chuX
chuW
chuT
chuA
chuS
iroB
iroC
iroD
iroE
iucA
iucB
iucC
iucD
iutA
kpsD
kpsE
kpsM
kpsT
1400
1600
1800
2000
1950
1970
1990
2010
1700
1800
1900
2000
time (calendar years)
O-groups
O1
O10
O101
O105
O107
O109
O11
O111
O117
O12
O124var1
O126
O129
O141
O148
O15
O153var1
O16
O17
O168
O18
O182
O2
O21
O25
O28ab
O29
O32
O40
O45
O71
O73
O8
O84
O88
O99
Onovel5
Onovel9
Unknown
H-groups
H10
H12
H18
H19
H21
H25
H27
H31
H32
H39
H4
H40
H48
H54
H55
H6
H7
H9
Unknown
FimH allele
FimH137
FimH15
FimH1515
FimH168
FimH23
FimH24
FimH25
FimH27
FimH279
FimH28
FimH30
FimH34
FimH41
FimH43
FimH47
FimH54
FimH566
FimH94
Gene content
resistance (absence)
resistance (presence)
virulence (absence)
virulence (presence)


A.
20
15
10
5
0
score
1980
2001
2010
year
VF category
adhesin
invasin
iron acquisition
miscellaneous
protectin
toxin
all VFs
B.
0.4
0.3
0.2
0.1
0.0
proportion
1980
2001
2010
year
Antibiotic category
aminoglycoside
betalactam
phenicol
quinolone
sulphonamide
tetracycline
trimethoprim
at least one antibiotic


A.
B.
ST driven change
(between ST change)
gene driven change
(within ST change)
ompT (chromosomal)
kpsD
kpsE
kpsM
capU
chu*
fyuA
irp2
iroB
iroN
iroC
usp
etsC
ibeA
cvaC
fdeC
iucA
iucB
iucC
iucD
papX
traT
iutA
espL1
espX4
kpsT
neuC
espX1
espL4
espY1
papB
focC
sfaB
tcpC
focF
focH
sfaD
senB_3
hlyA
cnf1
iss
gad
etsC
hek
espX1
cvaC
hlyF
overall increase
overall decrease
adhesin
invasin
iron acquisition
miscellaneous
protectin
toxin


A.
B.
frequency variation
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
Gene driven (1980-2001)
Gene driven (2001-2010)
ST driven (1980-2001)
ST driven (2001-2010)
VF category
adhesin
invasin
iron acquisition
miscellaneous
protectin
toxin
Antibiotic category
aminoglycoside
betalactam
macrolide
phenicol
sulphonamide
tetracycline
trimethoprim


A.
overall increase
overall decrease
aminoglycoside
betalactam
macrolide
phenicol
sulphonamide
tetracycline
trimethoprim
sul1
tet(A)
blaTEM-1B
ant(3'')-Ia
blaTEM-1A
dfrA1
aadA5
catA1
dfrA5
dfrA17
mph(A)
tet(B)
aph(6)-Id
aph(3'')-Ib
sul2
ST driven change
(between ST change)
gene driven change
(within ST change)
B.
aph(6)-Id
aph(3'')-Ib
aadA5
dfrA17
catA1
mph(A)
sul2
dfrA5
blaTEM(1C)
blaTEM(1B)
espY1
espY1
sul1
dfrA1
tet(A)