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

Julie Marin[1,2], Olivier Clermont[1,2], Guilhem Royer[1,2,3,4], Mélanie Mercier-Darty[5], Jean Winoc Decousser[4,6], Olivier Tenaillon[1,2], Erick Denamur[1,2,7*] and François Blanquart[1,2,8*].

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

[2]Université de Paris, INSERM, IAME, 75018 Paris, France

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

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

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

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

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

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

***: co-last authors**

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

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

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

ABSTRACT

Escherichia coli is a commensal species of the lower intestine, but also a major pathogen causing intestinal and extra-intestinal infections. Most studies on genomic evolution of *E. coli* used isolates from infections, and/or focused on antibiotic resistance, but neglected the evolution of virulence. Here instead, we whole-genome sequenced a collection of 436 *E. coli* isolated from fecal samples of healthy adult volunteers in France between 1980 and 2010. These isolates were distributed among 159 sequence types (STs), the five most frequent being ST10 (15.6%), ST73 (5.5%) and ST95 (4.8%), ST69 (3.7%) and ST59 (3.7%), and 230 O:H serotypes. ST and serotype diversity increased over time. Comparison with 912 *E. coli* bacteremia isolates from similar region and time showed a greater diversity in commensal isolates. The O1, O2, O6 and O25-groups used in bioconjugate O-antigen vaccine were found in only 63% of the four main STs associated with a high risk of bacteremia (ST69, ST73, ST95 and ST131). In commensals, STs associated with a high risk of bacteremia increased in frequency. Both extra-intestinal virulence-associated genes and resistance to antibiotics increased in frequency. Evolution of virulence genes was driven by both clonal expansion of STs with more virulence genes, and increases in frequency within STs, whereas the evolution of resistance was dominated by increases in frequency within STs. This study provides a unique picture of the phylogenomic evolution of *E. coli* in its human commensal habitat over a 30-year period and suggests that the efficacy of O-antigen vaccines would be threatened by serotype replacement.

KEY-WORDS: *Escherichia coli*, commensal, resistance, virulence, population genomics

INTRODUCTION

Escherichia coli is a commensal species of the lower intestine of humans and other vertebrates (Berg 1996; Tenaillon et al. 2010). It is also a major pathogen causing intestinal and extra-intestinal infections responsible for about a million deaths worldwide each year (Denamur et al. 2021; Kaper et al. 2004). Among the broad range of diseases caused by *E. coli*, the most common are urinary tract infections, bacteremia and intestinal infections (Russo and Johnson 2003; Kaper et al. 2004). It is the Gram-negative pathogen causing most community-acquired and hospital-acquired bacteremia (Goto et al. 2017), ranked third in the World Health Organization list of antibiotic resistant 'priority pathogens'.

E. coli strains share a core genome of about 2,000 genes (Touchon et al. 2009; Rasko et al. 2008). In addition, each strain has about 2,500-3,000 genes from the accessory genome, with the total pangenome containing more than tens of thousands of genes including virulence and antibiotic resistance genes (Rasko et al. 2008; Touchon et al. 2009, 2020). In spite of recombination involving double cross-over (Didelot et al. 2012), this species presents a robust clonal structure with at least nine phylogroups called A, B1, B2, C, D, E, F, G and H (Denamur et al. 2021), and sequence types (STs) defined at a finer genetic resolution. Phylogroups have distinct host specificities. As an example, extra-intestinal strains belong mainly to B2 and D phylogroups (Picard et al. 1999) whereas A and B1 are more generalist (Berthe et al. 2013; Power et al. 2005; Walk et al. 2007), being associated with all vertebrates and water environments (Touchon et al. 2020). Surface structures as O-polysaccharide and H-flagellar antigens (Orskov et al. 1977) as well as the *fimH* protein which is the receptor-recognizing element of type 1 fimbriae (Schembri et al. 2001), are key elements in the pathophysiological processes and epidemiological typing (Roer et al. 2017). O-antigens are used as a target for the development of *E. coli* vaccines to protect against extra-intestinal infections (Frenck et al. 2019; Huttner et al. 2017).

From a public health perspective, it is particularly important to understand the evolution of the repertoire of virulence and antibiotic resistance genes. Virulence genes are associated with a higher risk of urinary tract and bloodstream infections (Johnson 1991; Clermont et al. 2017). Antibiotic resistant infections are associated with longer hospital stay, increased risk of death and public health costs (Kraker et al. 2011). While temporal trends in virulence have been less often examined than in antibiotic resistance, virulence and antibiotic resistance genes repertoire greatly varies among strains, revealing continuous gene acquisition and loss (Didelot et al. 2009). The mean number of virulence genes in commensal *E. coli* has increased from 1980 to 2010 in France (Massot et al. 2016), although it is unclear if this is a local or more widespread trend. The incidence of *E. coli* bacteremia has increased in Europe, which could be explained by bacterial evolution towards higher virulence, but also by increased reporting, changing epidemiological factors like the ageing population or evolution of medical practices (Kraker et al. 2011; Vihta et al. 2018). In *E. coli,* resistances to multiple antibiotics have rapidly increased in frequency over the last decades, and stabilized at an intermediate level. For example, resistance to 3rd generation cephalosporins due mainly to CTX-M antibiotic-degrading enzymes has increased from the 1990s and seems to stabilize around 5% to 15% in commensal *E. coli* in Europe (Birgy et al. 2016; Woerther et al. 2013).

What evolutionary forces act on virulence and resistance genes? The first hypothesis is that these genes are direct targets of selection. Virulence genes are thought to primarily be selected in the intestine as a by-product of commensalism (Levin and Edén 1990; Le Gall et al. 2007; Diard et al. 2010). Virulence genes are linked with longer persistence in the gut (Nowrouzian et al. 2003; Östblom et al. 2011). Extra-intestinal compartments do not typically lead to efficient onward transmission and are often considered an evolutionary dead end (Diard et al. 2010). As a second hypothesis, clonal expansion could modify the prevalence of resistance and virulence genes, without these genes being the direct target of selection. Clonal expansion, when recombination is low, may be particularly important since several resistance and virulence genes are closely linked to specific STs (Manges et al. 2001; Nicolas-Chanoine et al. 2014; Martin et al. 2013). The selective forces acting on resistance are clearer. Resistance is selected by antibiotic use, as evidenced by the spatial correlation between local antibiotic use and resistance (Goossens et al. 2005; Low et al. 2019) and the fact that hosts are colonized by resistant strains more frequently after antibiotic use (Chatterjee et al. 2018). For *E. coli*, exposure to antibiotics results 95% of the time from antibiotic treatment prescribed for infections unrelated to *E. coli*, and not to treat an *E. coli* infection (Tedijanto et al. 2018). Such "bystander" antibiotic exposure happens around once per year in adults in France (Sabuncu et al. 2009).

Although the gut is the typical habitat of *E. coli* and likely the main ecological context of selection for virulence and resistance, most studies on the evolution of virulence and antibiotic resistance focused on pathogenic clinical strains isolated from extra-intestinal infections (de Lastours et al. 2020; Cole et al. 2019; Kallonen et al. 2017; Salipante et al. 2015). Extra-intestinal infections are a rare occurrence in *E. coli*'s life: the incidence of urinary tract infections is of the order of 10^{-2} per person-year (Foxman 2010), and of bacteremia of the order of 10^{-4} per person-year (Goto et al. 2017). In comparison to commensal *E. coli*, *E. coli* sampled from infections are biased towards strains carrying more virulent genes (Clermont et al. 2017; de Lastours et al. 2020; Kallonen et al. 2017). Strains from infections may also be less diverse than commensal strains as they belong to some specific clones (Denamur et al. 2021;

Kauffmann 1947), hindering the detection of temporal trends in these elements and the evolutionary mechanisms driving the emergence of pathogenic clones (Arimizu et al. 2019).

To avoid these biases, we investigated here the short-term evolution of commensal *E. coli* over 1980 to 2010 using strains from fecal samples of healthy volunteers in France. We whole-genome sequenced 436 strains gathered from several previous studies using the same methodology. We first examined the phylogenetic distribution of strains in phylogroups, STs and serotypes (O:H combinations) between 1980 and 2010. We compared the diversity of STs and O-groups in commensal *versus* pathogenic strains. Finally, we quantified the relative importance of ST clonal expansion and within-ST frequency change in explaining the recent evolution of virulence and resistance gene repertoires.

RESULTS

The pan-genome of the 436 strains contained 49,310 genes, with no saturation in the number of genes detected (figure S1). The core genome, defined as the number of genes found in more than 99% of the isolates, contained 2210 genes. A total of 431 genes were found in 95% to 99% of the isolates and 3281 genes were found in 15% to 95% of the isolates. The large majority of genes (43,388 out of 59,310) were found in less than 15% of the isolates.

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

Of the 436 isolates, 430 were assigned to the *E. coli* phylogroups A (131), B1 (66), B2 (125), C (10), D (51), E (13), F (27), G (6) and H (1) (tables S5-6). We found one strain belonging to the *Escherichia* clade III, two to the *Escherichia* clade IV and one to the *Escherichia* clade V. Two strains (C020-013 and LBC20a) were phylogenetically isolated from the known phylogroups (A to H). They probably belong to an undescribed phylogroup (figure S2). By fitting a binomial generalized linear model (GLM), we detected a significant decrease in frequency for the phylogroup A (from 58% in 1980 to 26% in 2010, p-value=1.55e-05) and a significant increase in frequency for the phylogroup B2 (from 9% in 1980 to 36% in 2010, p-value=9.96e-05) (figure 1A). No significant change was detected for the other phylogroups.

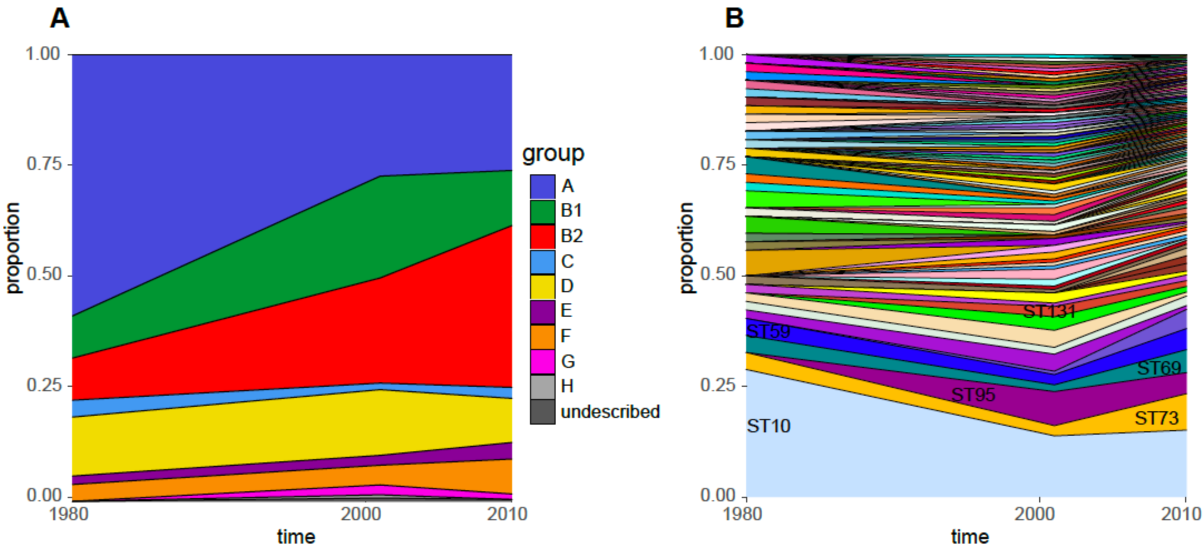

Figure 1. (A) Frequency distribution of phylogroups between 1980 and 2010. (B) Frequency distribution of STs between 1980 and 2010. The proportion of each ST has been plotted by year ordered by the overall frequency (most common at the bottom). Only the names of STs with an overall frequency > 0.3 and ST131 are shown. The only variation that was significant at the 0.05 level was the decline of ST10.

Those 436 isolates were distributed among 159 Warwick scheme STs (Wirth et al. 2006). The five most frequent STs were ST10 (15.6%) (phylogroup A), ST73 (5.5%) and ST95 (4.8%) (both of phylogroup B2), ST69 (3.7%) (phylogroup D) and ST59 (3.7%) (phylogroup F) (figure 1B and table 1). Using a binomial GLM, we detected a significant decrease in frequency for ST10 between 1980 and 2010 (from 28% in 1980 to 14% in 2010, p-value=0.021) but not for the four others most frequent STs. A total of 30, 73 and 97 distinct STs were detected in 1980, 2001 (2000-2002) and 2010 respectively. ST diversity was higher than expected in 2001, indicating an increase in ST diversity between 1980 and 2001 (figure 1B, figure S3A), independently from the changes in phylogroup frequencies (figure S3B).

Table 1. Distribution of the 11 more frequent sequence types (STs) of the *E. coli* commensal collection isolates (see Table S6 for the complete table). The number of isolates and the percentage for each year are presented in the table. The * indicate the STs for which we inferred the divergence times. We compared the ST diversity in a large collection of *E. coli* isolates from bacteremia in the Paris area (Royer et al. 2021) (isolates collected at years 2005 and 2016-2017 in approximately equal proportions) with our collection of commensal isolates in 2010, for all STs present in at least 5 strains in at least one of the two collections. We show the odds ratio (with 95% CI) for the risk of infection associated with colonization by each ST (logistic model of infection status as a function of the ST). Only the results for the 11 more frequent STs of the commensal collection are shown here, see Table S12 for the complete results. STs with odds ratio significantly different from 1 are highlighted in bold.

	Commensal				**Bacteremia**	**Odds ratio**
ST (phylogroup)	**1980**	**2001**	**2010**	**All years**	**2005 and 2016-2017**	
10* (A)	15 (28%)	18 (13%)	**35 (14%)**	68 (16%)	**34 (4%)**	**0.232 [0.141-0.382]**
73 (B2)	2 (4%)	3 (2%)	**19 (8%)**	24 (5%)	**110 (12%)**	**1.63 [1-2.79]**
95* (B2)	0 (0%)	10 (7%)	**11 (4%)**	21 (5%)	**84 (9%)**	**2.16 [1.18-4.34]**
69* (D)	2 (4%)	2 (1%)	**12 (5%)**	16 (4%)	**88 (10%)**	**2.07 [1.16-4.05]**
59 (F)	2 (4%)	3 (2%)	**11 (4%)**	16 (4%)	**5 (1%)**	**0.117 [0.0367-0.326]**
141 (B2)	0 (0%)	1 (1%)	**10 (4%)**	11 (4%)	**16 (2%)**	**0.42 [0.191-0.969]**
93 (A)	1 (2%)	5 (4%)	2 (1%)	8 (2%)	6 (1%)	0.805 [0.184-5.52]
452 (B2)	1 (2%)	2 (1%)	5 (2%)	8 (2%)	-	
405 (D)	1 (2%)	5 (4%)	2 (1%)	8 (2%)	10 (1%)	1.35 [0.352-8.8]
58 (B1)	0 (0%)	4 (3%)	3 (1%)	7 (2%)	27 (3%)	2.46 [0.86-10.4]
131 (B2)	0 (0%)	3 (2%)	**3 (1%)**	6 (1%)	**103 (11%)**	**10.3 [3.82-42]**
Total in collections	53	138	245	436	912	

We also detected an increase in serotype (O:H combination) diversity between 1980 and 2001, the serotype diversity being higher than expected in 2001 and 2010 (figure S4 and table S7).

However, the diversity of O-groups, H-types and *fimH* alleles remained stable between 1980 and 2010 when considered individually (figures S5-7 and tables S8-10).

We focused on the change in diversity of O and H antigens and *fimH* protein, widely involved in the pathophysiological process, within three major STs, ST10, ST69 and ST95. ST10 strains exhibit a large diversity of O and H antigens (44 serotypes) and *fimH* alleles (16 alleles). ST69 strains exhibit a limited diversity of O serogroups (O15, O17, O25b and O45) associated with H18 and *fimH*27 almost exclusively (figure 2). We found six serotypes (O:H antigens combinations) and five *fimH* alleles among ST95 strains. The apparent diversity in O:H combinations and *fimH* alleles reflects in part the age of the ST that can vary by a 20-fold factor according to our estimations (figure 2). The MRCA (most recent common ancestor) age of ST10, 1377 [866-1637], is much older than the MRCA age of ST69, 1951 [1924-1969], and ST95, 1680 [766-1882] (Bayesian skyline model, figure S8 and table S11). We tested whether the difference in their timescale of evolution explains the difference in diversity and found that the diversity of O and H antigens and the diversity of *fimH* increased faster for ST69 than for ST10 and ST95 (figure S9). We also quantified the change in diversity of epidemiological types (O:H serotype and *fimH* allele) within the most abundant ST of our commensal collection (ST10). We did not detect a significant change in frequency for any of the five most frequent serotypes within ST10 between 1980 and 2010 (figure S10).

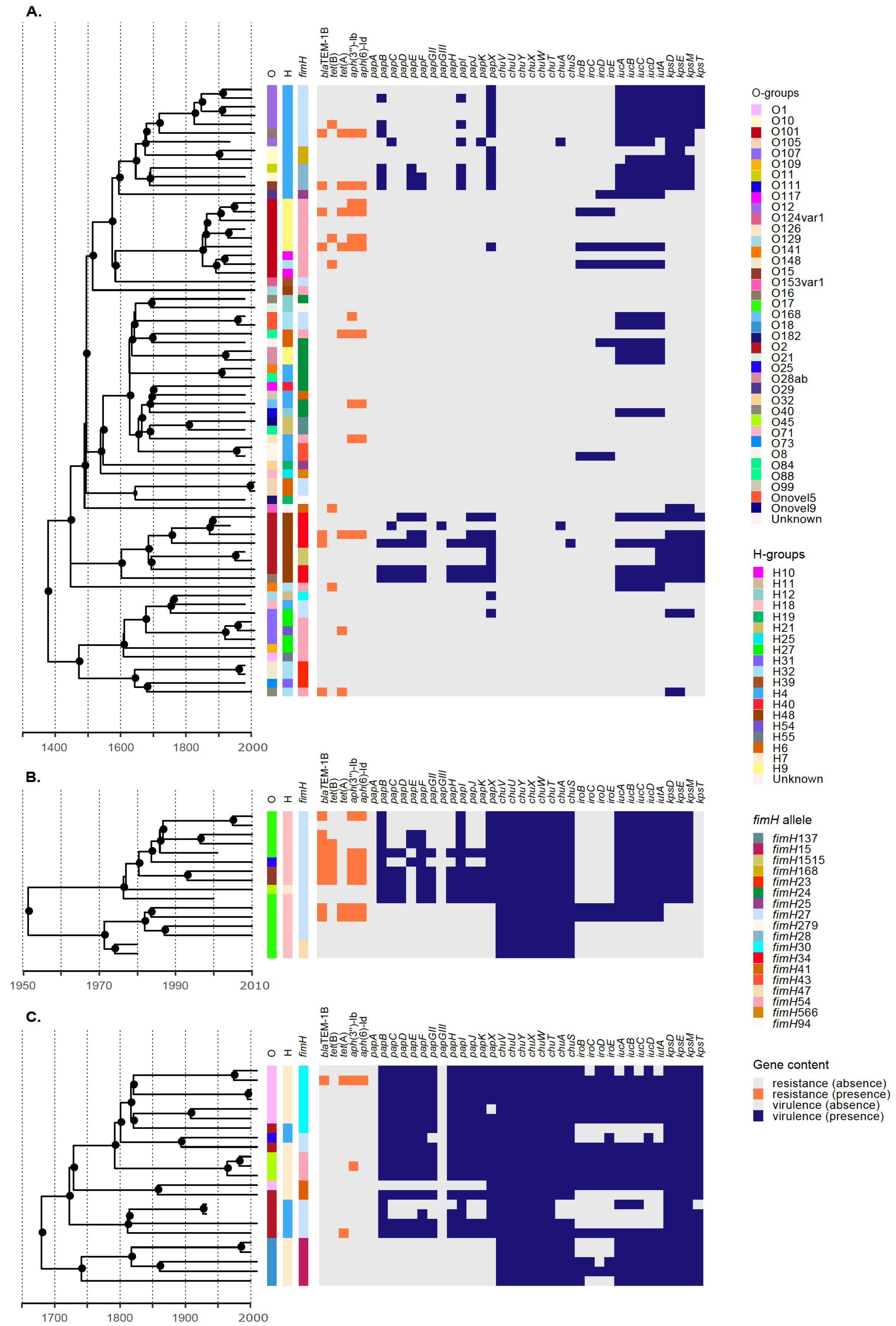

Figure 2. *(Caption opposite.)*

Figure 2. *(Opposite.)* Genomic content of ST10 (A), ST69 (B) and ST95 (C). In addition to the epidemiological type diversity (O:H serotypes and *fimH* allele), we examined the presence of five antibiotic resistance genes (*blaTEM-1B*, *tet(B)*, *tet(A)*, *aph(3'')-Ib* and *aph(6)-Id*) and of 34 virulence genes [*pap* (adhesin), *chu*, *iro*, *iuc* (iron capture systems) and *kps* (protectin) operons]. The timetrees were built with BEAST v1.10.4 (Drummond et al. 2012). ST10 exhibits the largest serotype diversity (44 serotypes), among the other STs examined here, ST69 (4 serotypes) and ST95 (6 serotypes). Nodes with a support value (Bayesian posterior probability) > 0.75 are indicated by black circles.

2. Diversity of STs and O-groups in commensal versus pathogenic strains

We compared our commensal collection with a set of 912 *E. coli* genomes collected from bacteremia at a similar location and date ("bacteremia collection") (Royer et al. 2021). The commensal collection was more diverse in its ST composition, with a higher number of rare STs and a lower number of frequent STs compared to the bacteremia collection (figure S11). The diversity of STs in commensal strains was very distinct to that in pathogenic strains (table 1 and table S12). Notably, S10 and ST59 are abundant in commensal strains (14.3% and 4.5% in 2010 respectively) but under-represented in bacteremia (3.7% and 0.5%); on the contrary, ST131, ST73, ST69, ST95 are less common in commensal strains than they are in bacteremia. ST95 and ST69 strains also had more virulence genes than ST10 (figure 2). This comparison can be translated in an odds ratio for the risk of infection associated with colonization by each ST, ST131 being the most pathogenic and ST59 the least pathogenic (table 1). Last, in our commensal collection, STs associated with a higher risk to cause infection in 2010 (odds ratio > 1) increased in frequency from 11% to 28% between 1980 and 2010 (figure S12).

The distribution of the O-group diversity also differed between the commensal and the bacteremia collections (table S13). The four O-groups targeted by the recently developed bioconjugate vaccine ExPEC4V (Frenck et al. 2019; Huttner et al. 2017), O1, O2, O6 and O25 are the most abundant O-groups in the bacteremia collection. However, O-groups O1, O2, O6 are not particularly associated with pathogenic strains (table S13). There is actually little association between pathogenicity and O-groups. Next, we evaluated the impact of the bioconjugate vaccine ExPEC4V on the four main STs associated with a higher risk of infection (ST69, ST73, ST95 and ST131) and on the three STs associated with a lower risk of infection (ST10, ST59 and ST141) by examining vaccine O-groups within these STs in the commensal collection. The vaccine would target 63% of strains within the most pathogenic STs (figure 3 and table S19) and 38% of the most commensal STs. Thus, the vaccine would not eliminate the most pathogenic STs as a sizable fraction of the STs escape the vaccine, and would affect the least pathogenic strains.

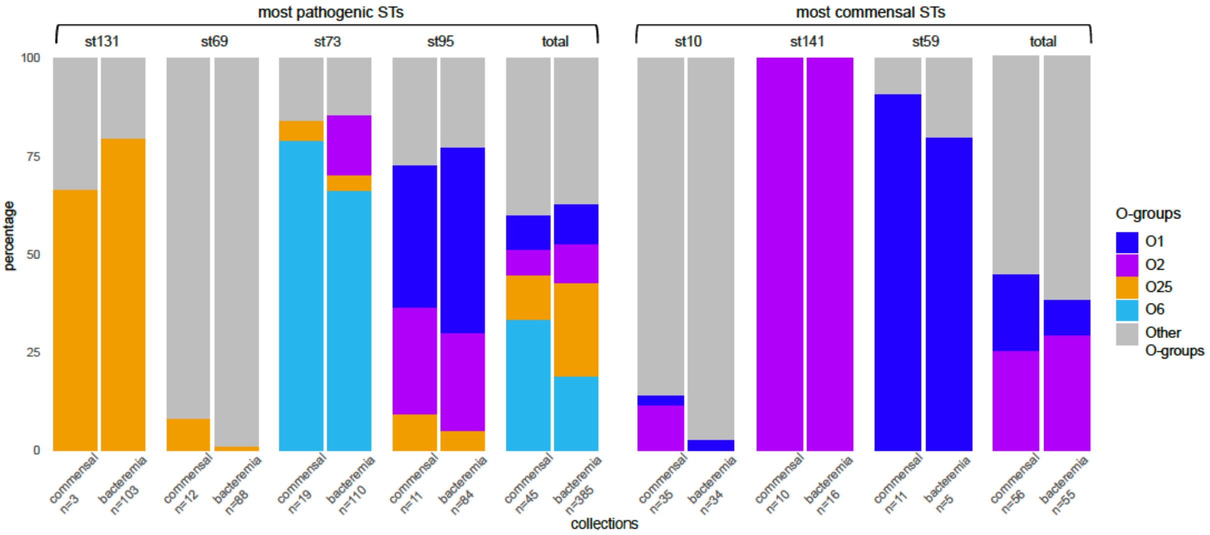

Figure 3. Distribution of the O-groups, O1, O2, O6 and O25, targeted by the ExPEC4V vaccine, within the four main STs associated with a high risk of infection and within the three main ST associated with a low of infection, in our commensal collection (year 2010) and in the bacteremia collection (Royer et al. 2021) (see Table S19). n: number of strains.

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

a. Recent increase in virulence and resistance

The correspondence analysis revealed three distinct groups of commensal *E. coli* isolates (figure S13). The first group, projected on the negative value of the first axis (Dim1), includes the phylogroups B2, F and G and H. These three phylogroups carry more VFs, in particular protectin and invasin genes, but fewer resistance genes. The second group, projected on the positive value of the first axis and negative values of the second axis, includes the phylogroups A, B1 and E. These three phylogroups carry more toxin genes. The third group, phylogroups C, D and the undescribed phylogroup, is projected on the positive value of both axes. These phylogroups are positively correlated with the large majority of resistances, including resistances to beta-lactam, sulphonamide, aminoglycoside, phenicol, tetracycline and trimethoprim antibiotics.

The prevalence of virulence increased between 1980 and 2010 (figure 4). First, the mean virulence score (number of virulence factors out of the 104 assayed) increased from 14 to 19. Using a linear model with sampling dates and phylogroups as predictors, we showed that, between 1980 and 2010, the increase in virulence score is explained by the change in the phylogenetic distribution; the effect of the sampling date is not significant (effect size = -8.42e-03 [-0.131; 0.114] VF product per year, p-value = 0.884). This increase is driven by the change in frequency of several phylogroups carrying many VFs, such as B2 (9% in 1980 to 36% in 2010 with a mean virulence score of 26, effect size = +12 [7; 18] compared to phylogroup A, p-value = 4.60-04) and F (4% in 1980 to 8% in 2010 with a mean virulence score of 25, effect size = +13 [7; 19] compared to phylogroup A, p-value = 3.58e-04). When testing each time period individually, we did not find a significant effect of the sampling date either, but only of the phylogenetic distribution (between 1980 and 2001: effect size = +0.053 [-0.128; 0.235] VF product per year, p-value = 0.484; between 2001 and 2010: effect size = -0.303 VF product per year [-0.852; 0.247], p-value = 0.226).

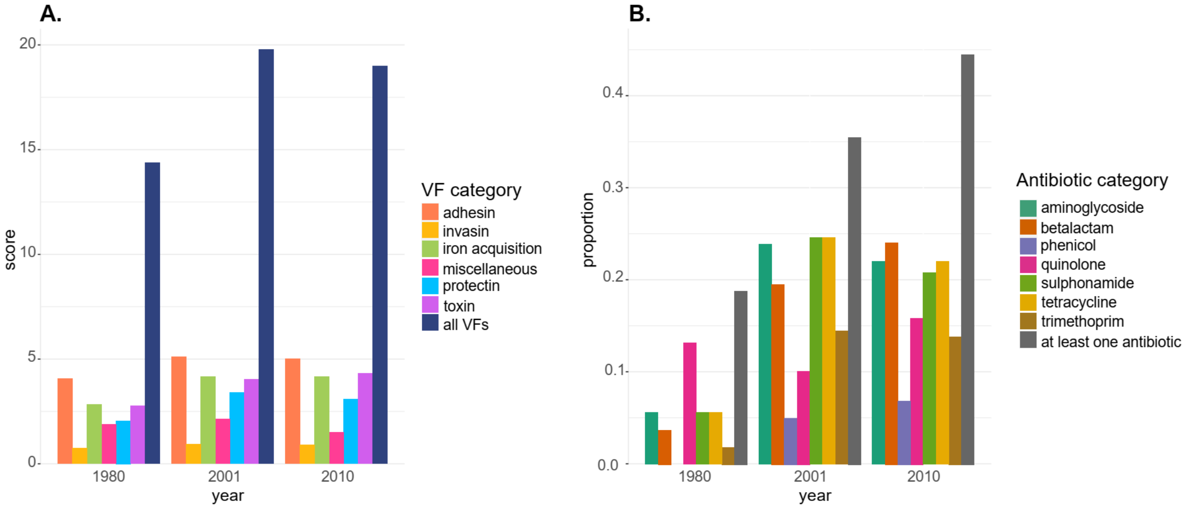

Figure 4. (A) Mean virulence score computed for all the 432 commensal *E. coli* isolates. The virulence score of a strain is defined as the number of the VF products of each category tested that were present in that strain. (B) Frequency of antibiotic resistance through time for all 432 isolates. Both gene acquisition and point mutations are included, but we omitted the macrolide category because more than 99% of the strains were resistant to macrolide antibiotics. For readability, in each panel the year 2001 includes strains sampled in 2000, 2001 and 2002.

The frequency of resistance also increased through time with the fraction of strains resistant to at least one antibiotic class increased from 19% to 44% from 1980 to 2010. This rise was significant within a linear model with sampling dates and phylogroups as predictors (effect size = +0.0113 [4.62e-03; 0.018] per year, p-value = 2.60e-03). In addition to sampling dates, this increase is also driven by the change in frequency of one minor phylogroup, phylogroup F which increases from 4% in 1980 to 8% in 2010 and for which 95% of the strains are resistant to at least one antibiotic (effect size = +0.687 [0.369; 1.00] compared to phylogroup A, p-value = 3.45e-04). When considering each time period individually, the frequency of resistance to one antibiotic significantly increased between 1980 and 2001 (effect size = +0.011 [1.16e-03; 0.020] per year, p-value = 0.033). The increase between 2001 to 2010 was comparable in size but not significant (effect size = +0.017 [-0.048; 0.038] per year, p-value = 0.111). Interestingly, the proportion of strains resistant to two or more antibiotic categories decreased between 2001 and 2010, from 30% to 22%.

b. ST clonal expansion *versus* increase in gene frequency within STs

We next evaluated how the temporal variation in the frequency of individual resistance and virulence genes was driven by phylogenetic constraints at the level of STs versus gene frequency changes independent of STs. STs are evolutionary units smaller than phylogroups but large enough to be followed through time. Nine STs are present in 1980, 2001 and 2010, they belong to the phylogroups A, B2, D and F and represent 36% of our data-set (table S14). We decomposed the temporal variation in gene frequency in two additive components. First, "ST-driven" change corresponds to a change in gene frequency driven by the change in frequency of STs (figure S14). Second, "gene-driven" change reflects the change in gene frequency within STs, independently of changes in ST frequencies. We therefore decomposed the evolution of virulence and resistance genes into "ST-driven" and "gene-driven" change (figures 5-7 and tables S15-18).

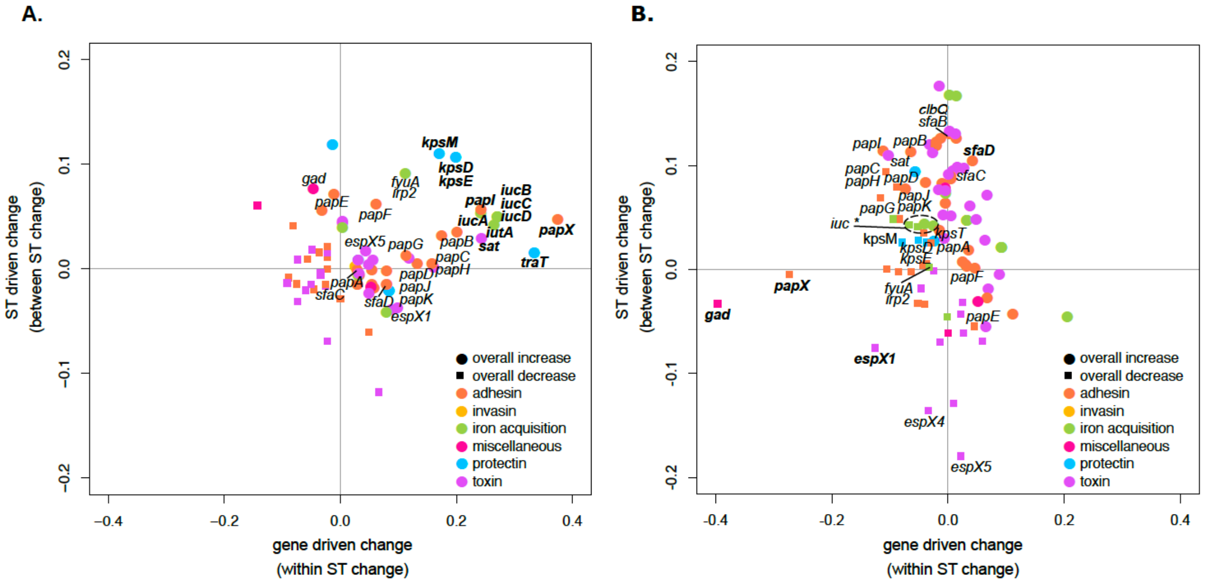

Figure 5. Temporal change of virulence frequency between 1980 – 2001 (A) and between 2001 – 2010 (B). The overall frequency change (Δf) for each gene (increases depicted by circles and decreases by squares) is decomposed in change driven by the variation in frequency of STs carrying the focal gene (ST driven change) and change driven by the variation in frequency of the focal gene (gene driven change). For readability, only genes for which between ST change or within ST change was greater than 0.02 in absolute value are shown (see table S15-S16 for the complete list). Genes highlighted in bold are those for which the temporal change is significant at the 0.05 level. In the panel B, *clbQ* and *sfaB* are superimposed and *iuc** includes *iucABCD* and *iutA*.

The evolution of virulence is both ST and gene driven. The relative contribution of the two processes depends on the time period and the genes considered. Frequency changes within STs (gene driven change) contributed mostly in the period between 1980 and 2001 (figures 5A and 6). Those changes are mainly driven by the large increase in frequency of the *kps* operon (K1 capsule, from 42% to 70-72%, p-values = 0.02 and 0.04 for the association between the gene presence and time of sample between 1980 and 2001 according to Fisher's exact test) and of the *iuc* operon (Aerobactin, from 31-35% to 60-65%, p-values = 0.02) (figure 5A and table S15). From 2001 to 2010, the change in virulence was driven by both the increase in frequency of more virulent STs (ST driven change) and the change in gene frequency within

STs (gene driven change). We also detected a relatively large increase in frequency of iron acquisition VFs (figure 6). The contribution of ST and gene driven changes also depend on the gene considered. The increase in frequency of the *iuc* operon, an iron acquisition system associated with bacteremia (Clermont et al. 2017) and death in a mouse model of sepsis (Galardini et al. 2020) which can be located on a plasmid, is mainly gene driven between 1980 and 2001. The gene *clbQ*, part of the colibactin gene cluster associated with human colorectal cancer (Dziubańska-Kusibab et al. 2020; Nougayrède et al. 2006) , increased in frequency from 10% in 2001 to 23% in 2010 (ST driven change) although this increase was not significant (table S16). The genes *fyuA* and *irp2*, part of the HPI ("High Pathogenicity Island") which is an iron acquisition system associated with bacteremia (Clermont et al. 2017; Schubert et al. 1998) and death in a mouse model of sepsis (Galardini et al. 2020), increased in frequency from 50% to 70% between 1980 and 2001 although this increase was not significant either (table S15).

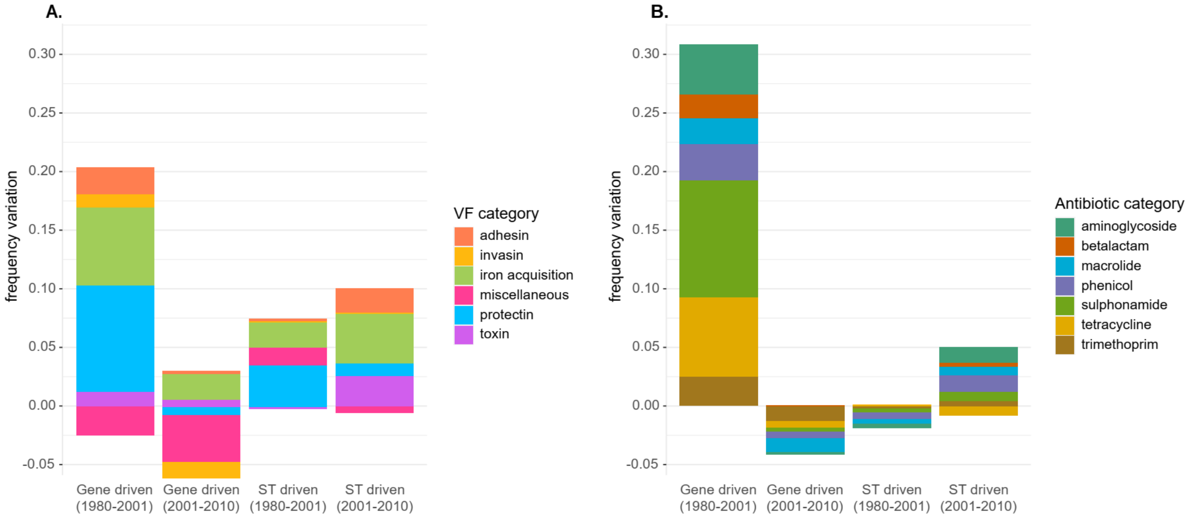

Figure 6. Summary of the decomposed temporal change of (A) virulence and of (B) antibiotic resistance between 1980 and 2010. For virulence, we computed the mean change per product (e.g. mean change among *iucA*, *iucB*, *iucC*, *iucD* and *iutA* genes for aerobactin), and reported the mean change among products of a category (e.g.

among products classified in the iron acquisition category). For resistance, we computed the mean change among genes belonging to a given category (e.g. among tet(A), tet(B), tet(D) and tet(M) for tetracycline).

In contrast to virulence, the evolution of resistance is almost exclusively gene driven and occurred primarily between 1980 and 2001 (figures 6-7 and tables S17-18). Significant increase in frequency was detected for genes responsible for resistance to beta-lactam (*bla*TEM-1B, from 0% to 17.5%, p-value = 0.037), tetracycline (*tet(A)*, from 0% to 17.5%, p-value = 0.037) and aminoglycoside (*aph(6)-Id*, from 0% to 20%, p-value = 0.018) antibiotics. Resistance genes frequency did not change much in the second period between 2001 and 2010. Thus, these genes rapidly increased in frequency in 20 years and subsequently stabilized, and their dynamics were unaffected by changes in the ST composition of the population.

Gene content visualization of ST10, ST69 and ST95 corroborate these results (figure 2). Virulence genes appear to be more phylogenetically clustered than resistance genes, explaining the importance of clonal expansion in driving the evolution of the virulence gene repertoire. We last tested if genes located on plasmids are more likely gene-driven (increasing simultaneously within multiple STs) than ST-driven (clonal expansion), but found no significant effect (correlation across genes between the proportion of gene driven change and plasmid *vs.* chromosome as a categorical variable).

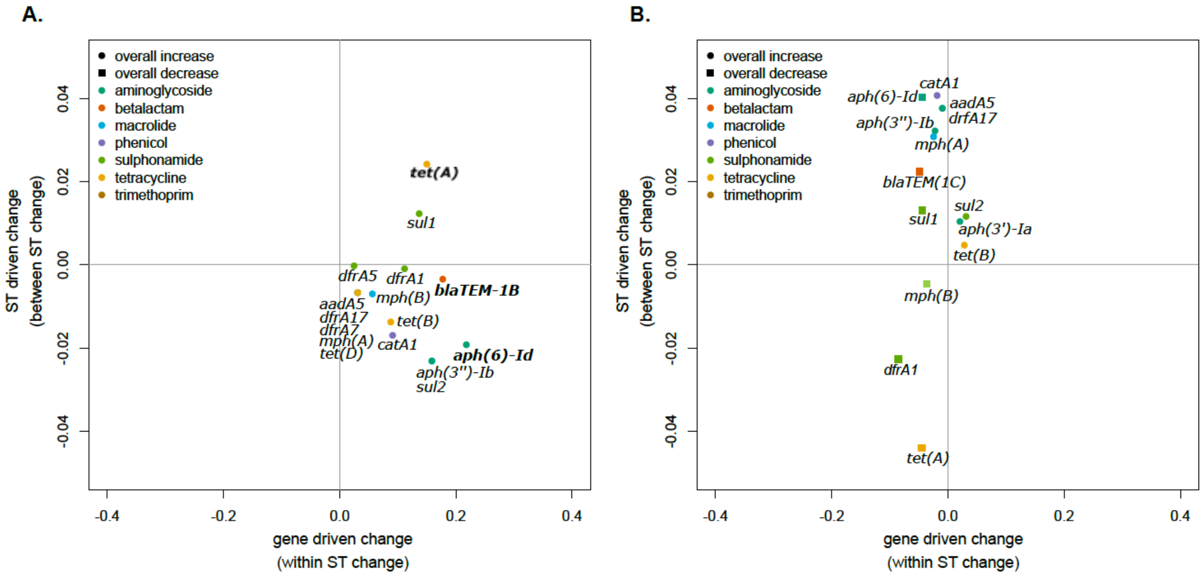

Figure 7. Temporal change of antibiotic resistance frequency between 1980 – 2001 (A) and between 2001 – 2010 (B). The overall frequency change (Δf) for each gene (increases depicted by circles and decreases by squares) is decomposed in change driven by the variation in frequency of STs carrying the focal gene (ST driven change) and change driven by the variation in frequency of the focal gene (gene driven change). For readability, only genes for which between ST change or within ST change was greater than 0.02 are shown here (see table S17-S18 for the complete list). Genes highlighted in bold are those for which the temporal change is significant at the 0.05 level. Note the scale of the y-axis is much smaller for resistance than for virulence (figure 5) as ST-driven changes are minor compared to gene-driven changes for antibiotic resistance genes.

DISCUSSION

1. The phylogenetic distribution of *E. coli* commensal strains

The primary habitat of *E. coli* is the gut. However, most of the studies interested in the study of virulence and resistance to antibiotics focused on pathogenic collections isolated from extra-intestinal infections, with a few rare exceptions (Bok et al. 2018; Qin et al. 2013; Raimondi et al. 2019; Smati et al. 2013; Arimizu et al. 2019). Here, using whole-genome analysis, we study the phylogenomic evolution of a large commensal collection of *E. coli* (436 strains) over a 30-year period.

The proportion of phylogroup B2 strains increased from 9% to 37% while the proportion of phylogroup A strains decreased from 58% to 26% between 1980 and 2010, as already observed in a large part of this collection (Massot et al. 2016). A predominance of B2 strains over other phylogroups in commensal samples has also been reported in different industrialized countries. For instance, in the late 1990s, the frequency of B2 strains in Australia, Japan, Sweden and USA is between 44 and 48% (Tenaillon 2010). It is possible that phylogroup B2 increased in frequency before the 1990s as they did in France from 1980 to 2010.

The diversity of ST was greater in 2001 and 2010 than in 1980. This was not explained by the increase in frequency of B2 strains (figure S3), but rather by the increase in frequency of rare STs from 1980 to 2001. Interestingly, several of the rare STs sampled since 2001 are associated with a higher risk to cause infections, such as ST14, ST58 and ST88 (figure S12).

Commensal strains were distinct from extra-intestinal pathogenic strains in ST composition. Differences in VF, O-group and phylogroup composition among commensal and extra-intestinal pathogenic strains have been often observed (e.g. Johnson et al. 2004; Kudinha et al. 2013; Clermont et al. 2017; Mereghetti et al. 2002; Kauffmann 1947) contrary to difference in clone diversity (Caugant et al. 1983). Here, we explicitly quantify the difference in ST diversity (ST number, size and distribution) among commensal and extra-intestinal pathogenic strains. Several STs, for example ST59, are over-represented in the commensal collection, while others like ST131 are over-represented in the bacteremia collection (table 1). Moreover, rare STs are more numerous in commensal than in pathogenic collections (figure S11). We also quantify the propensity of bacteria of each ST to escape from the gut and cause an extra-intestinal infection. The frequency of a ST in samples from extra-intestinal compartments is proportional to the product of the frequency of this ST in samples from the gut, and the propensity of the ST to cause an extra-intestinal infection. This disparity in composition justifies the interest in commensal collections, in addition to the more commonly studied extra-intestinal pathogenic collections, especially when the focus is the genetic diversity or the evolution of virulence and antibiotic resistance.

At a finer taxonomic scale, in addition to the clonal expansion of several epidemiological types (O:H serotype and *fimH* allele combinations), the homoplasy of several surface antigens and *fimH* alleles suggests a major role of horizontal transfers in shaping diversity of the three STs we studied in more detail ST10, ST95 and ST69 (figure 2). O-antigen coding and *fim* locus are major hot-spots of recombination in the genome (Touchon et al. 2009). Our focus on three major STs, two responsible for extra-intestinal infections in humans (ST69 and ST95) (Basmaci et al. 2015; Denamur et al. 2021) and one (ST10) found at high frequency in the human gut as well as in animals (Manges et al. 2015), revealed a larger diversity for ST10 as described elsewhere (Denamur et al. 2021; Royer et al. 2021). Diversification per time unit of O and H antigens and of *fimH* alleles for ST69 is faster than for ST10 and ST95, which its potentially explain by its younger age as it can be expected during adaptive radiations for example (Barrier et al. 2001).

Several O-groups have been considered as a promising target for a bioconjugate vaccine against extra-intestinal infections (Poolman and Wacker 2016). Recently, a phase 2 randomized controlled trial showed that a vaccine targeting the O1, O2, O6, and O25-antigens was well tolerated and elicited an antibody response against these antigens (Frenck et al. 2019; Huttner et al. 2017). However, despite being the most abundant O-groups, these four O-groups are not at risk of infection when considering the odds ratio, with the exception of the O25 (table S13). This could lead to clonal replacement by other non-vaccine pathogenic clones, as observed for the 13-valent pneumococcal conjugate vaccine (PCV13) (Ouldali et al.

2021), questioning the long-term efficacy of this vaccine. Furthermore, such vaccine could also perturb the commensal gut microbiota by eliminating, in addition to the major pathogenic clones, commensal *E. coli* clones (figure 3).

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

Gene and phylogroup frequencies varied in time (figure 1 and 4) (Jauréguy et al. 2007; Massot et al. 2016; Touchon et al. 2020; Escobar-Páramo et al. 2004a). Both B2 strains, which carry many VFs, and VFs increased in frequency, whereas the observed increase in resistance seems decoupled from the stability in frequency of B1 and C strains, associated to antibiotic resistance genes. To investigate whether virulence and antibiotic resistance evolution are governed by different evolutionary mechanisms, we decomposed the change in frequency of a gene clonal expansion of STs carrying this gene and increase in gene frequency within STs.

The frequency of STs can vary in time, randomly or in response to selective pressures. As a consequence, the frequency of a focal gene varies accordingly to the variation in frequency of the STs carrying it (whether or not the gene is under selection). The change in gene frequency may result from vertical transfers (clonal expansion) and horizontal transfers within STs. We called this process ST driven change. The frequency of a gene can also vary concomitantly in one or several STs regardless of variation in ST frequencies. This can result from horizontal transfers between or within STs or the increase in frequency of the lineage carrying the focal gene at the expense of others within STs. We called this process gene driven change. We partitioned the overall change in gene frequency in two terms, ST driven change and gene driven change, to assess the relative contribution of each process.

The contributions of gene driven and ST driven change varied between virulence and resistance genes, and between the two periods (1980-2000 and 2000-2010) (figure 5-7).

Between 1980 and 2000, the virulence of *E. coli* strains rapidly evolved through both the rise of more virulent STs and within-ST increase in frequency of virulence genes (gene driven change) (figures 5A and 6A). In the second time-period from 2001 to 2010, the increase in frequency of STs was mainly driven by the increase in frequency of more virulent STs (ST driven change) (figures 5B and 6A). Significant within-ST changes in frequency suggest a role for direct selection on virulence genes in driving the increased virulence. However, we cannot exclude that virulent STs increased in frequency as a result of other processes independent of selection on virulence genes.

What factors could explain the recent increase in virulence? Recent environmental changes include a shift towards more processed, and more nutrient-dense food over the period 1969-2002 in France, and a stabilization over 2002-2010 (Caillavet et al. 2018), mirroring the spread of virulence genes within ST observed in the period 1980-2000. This could directly select for virulence genes (e.g. iron acquisition factors) or indirectly select for virulent STs. Whether bacteria carrying some virulence genes are better adapted to nutrient-dense diets could be experimentally tested (O'Brien and Gordon 2011). In an effort to do so, we recently analyzed how mice diet affected the density of *E. coli* in the mice gut and found that the B2 strains used was more prevalent in a high sugar high fat diet than in a high fiber diet (Ghalayini et al. 2019). In addition to nutriment availability, protection again host immune defense, phages and protozoans also drives the evolution of virulence (Denamur et al. 2021; Wildschutte et al. 2004). A change in the diversity and abundance of these predators might likewise explain the recent increase in virulence observed in commensal *E. coli*.

For resistance genes, the rapid increase in frequency of several genes [notably *bla*$_{\text{TEM-1B}}$, *tet*(A), and *aph*(6)-Id which are located on plasmid or transposons (Heffron et al. 1975; Khezri et al. 2021; Ribera et al. 2003)] from 1980 to 2000 was mainly driven by an increase in the frequency of these genes independently within several STs (figure 7A). In contrast to 1980-2000, after 2001 nearly no change in frequency for resistance was detected (figures 6B-7). In the early 2000s, levels of antibiotic resistance were higher in France than in most European countries (Sabuncu et al. 2009; Goossens et al. 2005). A nationwide awareness campaign was launched in 2001 leading to a -26.5% reduction in antibiotic use in humans over 5 years (Sabuncu et al. 2009). The reduction in antibiotic use could explain the stabilization in resistance gene frequencies observed in our data.

Our work has some limitations. First, only one isolate per subject was sampled. It is well known that populations of subdominant clones are present in the feces (Smati et al. 2013). Resistant clones are often subdominant and isolated using antibiotic containing plates. They have a specific population structure and epidemiology (Day et al. 2019). Interestingly, nine of 10 most prevalent STs in clinical ESBL-producing *E. coli* were also the most common types in community feces, the exception being ST131 which was rare in our dataset (Verschuuren et al. 2021). Second, our data set ends in 2010 and we did not capture the recent evolution. Nevertheless, our collection of strains gathered in the same conditions over a 30-year period in a single location (France) represents a unique material.

CONCLUSION

To investigate the evolutionary mechanisms responsible for the recent increase in virulence and antibiotic resistance observed over 30 years in France, we whole-genome sequenced a large collection of 436 dominant commensal *E. coli* sampled from the gut of healthy volunteers. Commensal strains are more diverse than extra-intestinal pathogenic strains and distinct in their sequence type and serotype composition. Several antibiotic resistance genes rapidly spread from 1980 to 2000, largely unhindered by clonal structure. Higher virulence evolved through increase in virulence gene frequency within STs and clonal expansion of more virulent STs. Increasing virulence of *E. coli* would result, everything else being equal, in an increasing incidence of extra-intestinal infections, and could indeed contribute to the increasing incidence of bacteremia observed in the last decades. Future research should investigate whether the observed increasing virulence in commensal strains was observed in other geographical areas and what environmental factors could select for *E. coli* virulence. Lastly, given the diversity of O antigens in commensal strains, the efficacy of vaccines against *E. coli* extra-intestinal diseases would be threatened by replacement with non-vaccine serotypes that are as pathogenic as vaccine serotypes.

METHODS

1. Strain collections

We studied the whole genomes of four hundred and thirty-six *E. coli* strains gathered from stools of 436 healthy adults living in the Paris area or Brittany (both locations in the North of France) between 1980 to 2010. These strains come from five previously published collections: VDG sampled in 1980 (Duriez et al. 2001), ROAR in 2000 (Skurnik et al. 2016), LBC in 2001 (Escobar-Páramo et al. 2004b), PAR in 2002 (Escobar-Páramo et al. 2004b) and Coliville in 2010 (Massot et al. 2016) (table S1). In all study, one single *E. coli* colony randomly picked on the Drigalski plate was retained per individual (Massot et al. 2016), representing probably the dominant strain. The study was approved by the ethics evaluation committee of Institut National de la Santé et de la Recherche Médicale (INSERM) (CCTIRS no. 09.243, CNIL no. 909277, and CQI no. 01-014).

In order to improve the temporal sampling, we used 24 sequences from the Murray collection with samples ranging from 1930 to 1941 (Baker et al. 2015) (table S2). From the 50 genomic sequences available in the Murray collection, we selected sequences with available sampling time and excluded multiple variants per strain (when strain name, date on tube and origin were identical for two samples).

2. Sequencing of the commensal strains

After DNA extraction, whole-genome of each strain was sequenced using Illumina NextSeq 2x150 bp after NextEra XT library preparation (Illumina, San Diego, CA) as in de Lastours et al. (de Lastours et al. 2020).

3. Assembly and typing

The assembly of genomes was performed using the in-house script petanc that integrates several existing bacterial genomic tools (Bourrel et al. 2019), including SPADES (Bankevich et al. 2012). This in-house script was also used to perform the typing of strains with several genotyping schemes using the genomic tool SRST2 (Inouye et al. 2014). Multilocus sequence typing (MLST) was performed and STs were defined using the Warwick MLST scheme (Wirth et al. 2006) and the Pasteur scheme (Jaureguy et al. 2008). We also determined the O:H serotypes (Ingle et al. 2016) and the *fimH* alleles (Roer et al. 2017). The phylogroups were defined using the ClermonTyping method (Beghain et al. 2018).

4. Distribution of phylogroups and STs through time

Because there were three time points close in time with a small number of strains, we aggregated data of years 2000, 2001 and 2002 when generating the stacked area charts of phylogroups and STs.

The number of strains analyzed varied among years, 53 in 1980, 138 between 2000 and 2002, and 245 in 2010. To determine whether the number of distinct STs detected each year varied through time or was simply reflecting the number of sampled strains, we generated null distributions of ST number depending on the sampling effort. We sampled 10,000 times a fixed number of strains (53 for 1980, 138 for 2001 and 245 for 2010) with the frequency of each ST set to its overall frequency. For each time point, the observed value was compared to the corresponding null distribution.

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

5. Comparison of commensal and previous pathogenic *E. coli* collection

We evaluated the risk of infection associated to colonization by a specific ST and by a specific O-group. We compared the ST and O-group diversity from a collection of 912 bacteremia isolates (isolates collected at years 2005 and 2016-2017 in approximately equal proportions) (Royer et al. 2021) with the most recent isolates of our commensal isolates (2010), for all STs with at least 5 strains in at least one of the two collections and for all O-groups with at least 5 strains in at least one of the two collections. The odds ratios for the infection risk were computed by fitting a logistic model of infection status (commensal or bacteremia) as a function of the ST or the O-group (here and thereafter, “significant” refers to significance at the 0.05 level).

Next, we compared the phylogenetic distribution of our commensal collection with the bacteremia collection. We calculated the cumulative frequency distribution of STs in the commensal collection, and we compared it to the same distribution in 200 random sub-samples of 436 sequences from the bacteremia collection.

Finally, we evaluated the distribution of four O-groups, O1, O2, O6 and O25, used in the ExPEC4V bioconjugate vaccine (Frenck et al. 2019), within the four main STs associated to a higher risk of infection (ST69, ST73, ST95 and ST131) in the most recent isolates of our commensal collection (2010) and in the bacteremia isolates collection (2005 and 2016-2017) (Royer et al. 2021).

6. Genomic diversity of the core genome

The 436 assemblies (our commensal collection) were annotated with Prokka (Seemann 2014). We then performed pan-genome analysis from annotated assemblies with Roary (Page et al. 2015) on the 436 genomes using default parameters. The alignment of the core genome and the list of genes of the accessory genome were generated.

7. Core genome phylogeny of *E. coli*

To build the phylogeny of *E. coli*, we aligned whole genomes (436 strains) to the reference R1B5J10 with Snippy 4.4.0 using standard parameters (Seemann 2015). We did not remove recombination events because it could accentuate errors in phylogenetic distances and the topology of the tree is usually not affected (Hedge and Wilson 2014; Lapierre et al. 2016). The alignment of 773,466 SNPs (single-nucleotide polymorphism) obtained with SNP-sites (Page et al. 2016) was used to produce a maximum likelihood phylogenetic tree with with RAxML (Stamatakis 2014) using the GTRGAMMA model with 1,000 bootstrap replicates.

8. Divergence times estimates for major STs

To further study the five most prevalent STs, ST10, ST73, ST95, ST69 and ST59, we aligned whole genomes (145 strains from our collection and 10 from the Murray collection) to references, R1B5J10, 016-002, R1B6J15, H1-004-0023-R-J and IAI39 respectively with Snippy 4.4.0 using standard parameters (Seemann 2015). For each ST, three outgroup sequences were selected as the three closest sequences to the focal ST belonging to two distinct and well supported clades, from the core genome phylogenetic tree. Recombination events were excluded with Gubbins using default settings (Croucher et al. 2015). The five resulting alignments were used to produce a maximum likelihood phylogenetic tree with RAxML (Stamatakis 2014) using the GTRGAMMA model with 1,000 bootstrap replicates.

The temporal analysis was performed with the program BEAST v1.10.4 (Drummond et al. 2012), on three of the most frequent STs with a molecular clock signal (i.e. positive relationship between root-to-tip distance and time): ST10, ST95 and ST69. BEAST implements the “uncorrelated relaxed clock” to model uncorrelated rate variations among lineages; the evolutionary rate of each branch is drawn independently from a common underlying distribution. To estimate divergence times with this model we used the following parameters: uncorrelated log-normal clock, substitution model GTR+I+G, fixed topology (RAxML tree) and 500 million generations with sampling every 1000 generations and a burn-in of 50 million generations. We ran three replicates with three tree priors: coalescent with constant population, coalescent with exponential growth and coalescent with Bayesian skyline. The sample times were used to calibrate the tips. For each data set, the best fitting model set was determined by computing Bayes factors from marginal likelihoods estimations calculated by stepping-stone sampling (Baele et al. 2013). Only the models that converged well and had and effective sample size (ESS) larger than 200 for each parameter were compared. The best fitting model was used in the subsequent analyses.

9. Virulence and resistance gene repertoire analyses

The resistome, the virulome and the plasmid type were defined using the in-house script petanc (Bourrel et al. 2019). They were defined by BlastN with Abricate (https:// github.com/tseemann/abricate) using the ResFinder database (Zankari et al. 2012), a custom database including the VirulenceFinder database (Joensen et al. 2014) and VFDB (Chen et al. 2016) to which we added selected genes (table S3), and PlasmidFinder respectively (Carattoli et al. 2014). We also searched for point mutation responsible for betalactam (*ampC* promoter) and fluroquinolone (*gyrA*/*B*, *parC*/*E*) resistance {Citation}. We set the threshold for minimum identity to 80% and for minimum coverage to 90%. If several copies of a gene were detected for a strain we only kept the copy with the maximal sum of coverage and identity. The plasmid sequences were also predicted by PlaScope (Royer et al. 2018).

To visually explore the relationships between the phylogroups and the resistance and virulence gene repertory content we performed a correspondence analysis (CA) (Benzecri 1992) with the R package 'FactoMineR' (Lê et al. 2008). A CA is a multivariate graphical analysis used to explore the associations among categorical variables. For each item in a table, a set of factor scores (coordinates) is obtained from linear combinations of rows and columns. These coordinates are projected on two dimensions, with in our case, the first axis opposing the phylogroups with the largest differences. The further the gene categories are from the origin the more they are discriminating. If two variables are close to each other in the plan, they are considered to be strongly associated.

For each isolate, we computed a virulence score corresponding to the number of the virulence factors (VF) present in each isolate (number of products out of the 104 tested, table S3) (adapted from Lefort et al. (2011)). The frequency of antibiotic resistance was computed for each year, both gene acquisitions and point mutations were included (table S4). The corresponding antibiotics were retrieved from ResFinder (Bortolaia et al. 2020). We next tested whether the virulence score and the frequency of antibiotic resistance changed in time. We fitted a linear model for the number of VF products or for the frequency of antibiotic resistance as a function of year (1980, 2001 (2000-2002), 2010) as a categorical variable.

To decipher whether changes in gene frequency through time were due to variations in ST frequencies or to variations in gene frequencies within STs, we decomposed the overall change in gene frequency (Δf) as follows (figure S14). We call $f_{i,t}$ the frequency of the focal gene in ST *i* at time *t.* We call $p_{i,t}$ the frequency of the ST *i* at time *t*. We are interested in the change in frequency of the focal gene from time t_1 to time t_2. The change in gene frequency can be decomposed as such:

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

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* 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 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 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.

For virulence, when the frequency of several alleles for a single gene was assessed by our in-house script, we combined those frequencies for each corresponding gene (table S3).

For each time period, and for each gene, we used Fisher's exact test to examine the significance of the association between gene presence/absence and time period.

The decomposed temporal changes of virulence and of antibiotic resistance were summarized as follow. For virulence, we first computed the mean change per product, and then the mean change among product for each category (see table S3). For resistance, we computed the mean change among genes for each category (see table S4).

DATA ACCESS

The data generated in this study have been submitted to the NCBI BioProject database under the accession numbers PRJEB39252 (Coliville), PRJEB38489 (ROAR), PRJEB44819 (LBC), PRJEB44872 (PAR), PRJEB44873 (VDG).

COMPETING INTEREST STATEMENT

The authors did not report a relevant conflict of interest.

ACKNOWLEDGMENTS

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing resources. Julie Marin and François Blanquart 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.

REFERENCES

Arimizu Y, Kirino Y, Sato MP, Uno K, Sato T, Gotoh Y, Auvray F, Brugere H, Oswald E, Mainil JG, et al. 2019. Large-scale genome analysis of bovine commensal *Escherichia coli* reveals that bovine-adapted *E. coli* lineages are serving as evolutionary sources of the emergence of human intestinal pathogenic strains. *Genome Res* **29**: 1495–1505.

Baker KS, Burnett E, McGregor H, Deheer-Graham A, Boinett C, Langridge GC, Wailan AM, Cain AK, Thomson NR, Russell JE, et al. 2015. The Murray collection of pre-antibiotic era Enterobacteriacae: a unique research resource. *Genome Med* **7**: 97.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.

Barrier M, Robichaux RH, Purugganan MD. 2001. Accelerated regulatory gene evolution in an adaptive radiation. *Proc Natl Acad Sci* **98**: 10208–10213.

Basmaci R, Bonacorsi S, Bidet P, Biran V, Aujard Y, Bingen E, Béchet S, Cohen R, Levy C. 2015. *Escherichia coli* meningitis features in 325 children from 2001 to 2013 in France. *Clin Infect Dis* **61**: 779–786.

Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. 2018. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb Genomics* **4**.

Benzecri J. 1992. *Correspondence analysis handbook.* Marcel Decker, New York.

Berg RD. 1996. The indigenous gastrointestinal microflora. *Trends Microbiol* **4**: 430–435.

Berthe T, Ratajczak M, Clermont O, Denamur E, Petit F. 2013. Evidence for coexistence of distinct *Escherichia coli* populations in various aquatic environments and their survival in estuary water. *Appl Environ Microbiol* **79**: 4684–4693.

Birgy A, Levy C, Bidet P, Thollot F, Derkx V, Béchet S, Mariani-Kurkdjian P, Cohen R, Bonacorsi S. 2016. ESBL-producing *Escherichia coli* ST131 versus non-ST131: evolution and risk factors of carriage among French children in the community between 2010 and 2015. *J Antimicrob Chemother* **71**: 2949–2956.

Bok E, Mazurek J, Myc A, Stosik M, Wojciech M, Baldy-Chudzik K. 2018. Comparison of Commensal *Escherichia coli* isolates from adults and young children in Lubuskie province, Poland: virulence potential, phylogeny and antimicrobial resistance. *Int J Environ Res Public Health* **15**.

Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, et al. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* **75**: 3491–3500.

Bourrel AS, Poirel L, Royer G, Darty M, Vuillemin X, Kieffer N, Clermont O, Denamur E, Nordmann P, Decousser J-W, et al. 2019. Colistin resistance in Parisian inpatient faecal *Escherichia coli* as the result of two distinct evolutionary pathways. *J Antimicrob Chemother* **74**: 1521–1530.

Caillavet F, Darmon N, Létoile F, Nichèle V. 2018. Is nutritional quality of food-at-home purchases improving? 1969–2010: 40 years of household consumption surveys in France. *Eur J Clin Nutr* **72**: 220–227.

Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. *In Silico* Detection and typing of plasmids using

PlasmidFinder and Plasmid Multilocus sequence typing. *Antimicrob Agents Chemother* **58**: 3895–3903.

Caugant DA, Levin BR, Lidin-Janson G, Whittam TS, Edén S, Selander RK. 1983. Genetic diversity and relationships among strains of *Escherichia coli* in the intestine and those causing urinary tract infections. *Host Parasite Relatsh Gram-Negat Infect* **33**: 203–227.

Chatterjee A, Modarai M, Naylor NR, Boyd SE, Atun R, Barlow J, Holmes AH, Johnson A, Robotham JV. 2018. Quantifying drivers of antibiotic resistance in humans: a systematic review. *Lancet Infect Dis* **18**: e368–e378.

Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res* **44**: D694-697.

Clermont O, Couffignal C, Blanco J, Mentré F, Picard B, Denamur E, Groups the C and C. 2017. Two levels of specialization in bacteraemic *Escherichia coli* strains revealed by their comparison with commensal strains. *Epidemiol Infect* **145**: 872–882.

Cole BK, Ilikj M, McCloskey CB, Chavez-Bueno S. 2019. Antibiotic resistance and molecular characterization of bacteremia *Escherichia coli* isolates from newborns in the United States. *PLOS ONE* **14**: e0219352.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**: e15–e15.

Day MJ, Hopkins KL, Wareham DW, Toleman MA, Elviss N, Randall L, Teale C, Cleary P, Wiuff C, Doumith M, et al. 2019. Extended-spectrum β-lactamase-producing

Escherichia coli in human-derived and foodchain-derived samples from England, Wales, and Scotland: an epidemiological surveillance and typing study. *Lancet Infect Dis* **19**: 1325–1335.

de Lastours V, Laouénan C, Royer G, Carbonnelle E, Lepeule R, Esposito-Farèse M, Clermont O, Duval X, Fantin B, Mentré F, et al. 2020. Mortality in *Escherichia coli* bloodstream infections: antibiotic resistance still does not make it. *J Antimicrob Chemother* **75**: 2334–2343.

Denamur E, Clermont O, Bonacorsi S, Gordon D. 2020. The population genetics of pathogenic *Escherichia coli*. *Nat Rev Microbiol* **19**: 37–54.

Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I. 2010. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. *J Bacteriol* **192**: 4885–4893.

Didelot X, Darling A, Falush D. 2009. Inferring genomic flux in bacteria. *Genome Res* **19**: 306–317.

Didelot X, Méric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13**: 256.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**: 1969–1973.

Duriez P, Clermont O, Bonacorsi S, Bingen E, Chaventré A, Elion J, Picard B, Denamur E. 2001. Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology,* **147**: 1671–1676.

Dziubańska-Kusibab PJ, Berger H, Battistini F, Bouwman BA, Iftekhar A, Katainen R, Cajuso T, Crosetto N, Orozco M, Aaltonen LA. 2020. Colibactin DNA-damage signature indicates mutational impact in colorectal cancer. *Nat Med* 1–7.

Escobar-Páramo P, Clermont O, Blanc-Potard A-B, Bui H, Le Bouguénec C, Denamur E. 2004a. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol Biol Evol* **21**: 1085–1094.

Escobar-Páramo P, Grenet K, Menac'h AL, Rode L, Salgado E, Amorin C, Gouriou S, Picard B, Rahimy MC, Andremont A, et al. 2004b. Large-scale population structure of human commensal *Escherichia coli* isolates. *Appl Environ Microbiol* **70**: 5698–5700.

Foxman B. 2010. The epidemiology of urinary tract infection. *Nat Rev Urol* **7**: 653–660.

Frenck RW, Ervin J, Chu L, Abbanat D, Spiessens B, Go O, Haazen W, van den Dobbelsteen G, Poolman J, Thoelen S, et al. 2019. Safety and immunogenicity of a vaccine for extra-intestinal pathogenic *Escherichia coli* (ESTELLA): a phase 2 randomised controlled trial. *Lancet Infect Dis* **19**: 631–640.

Galardini M, Clermont O, Baron A, Busby B, Dion S, Schubert S, Beltrao P, Denamur E. 2020. Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus *Escherichia* revealed by a genome-wide association study. *PLOS Genet* **16**: e1009065.

Ghalayini M, Magnan M, Dion S, Zatout O, Bourguignon L, Tenaillon O, Lescat M. 2019. Long-term evolution of the natural isolate of *Escherichia coli* 536 in the mouse gut colonized after maternal transmission reveals convergence in the constitutive expression of the lactose operon. *Mol Ecol* **28**: 4470–4485.

Goossens H, Ferech M, Vander Stichele R, Elseviers M. 2005. Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *The Lancet* **365**: 579–587.

Goto M, McDanel JS, Jones MM, Livorsi DJ, Ohl ME, Beck BF, Richardson KK, Alexander B, Perencevich EN. 2017. Antimicrobial nonsusceptibility of gram-negative bloodstream isolates, Veterans Health Administration System, United States, 2003–20131. *Emerg Infect Dis* **23**: 1815–1825.

Hedge J, Wilson DJ. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *mBio* **5**. https://mbio.asm.org/content/5/6/e02158-14 (Accessed May 21, 2021).

Heffron F, Sublett R, Hedges RW, Jacob A, Falkow S. 1975. Origin of the TEM-beta-lactamase gene found on plasmids. *J Bacteriol* **122**: 250–256.

Huttner A, Hatz C, van den Dobbelsteen G, Abbanat D, Hornacek A, Frölich R, Dreyer AM, Martin P, Davies T, Fae K, et al. 2017. Safety, immunogenicity, and preliminary clinical efficacy of a vaccine against extraintestinal pathogenic *Escherichia coli* in women with a history of recurrent urinary tract infection: a randomised, single-blind, placebo-controlled phase 1b trial. *Lancet Infect Dis* **17**: 528–537.

Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, Levine MM, Robins-Browne RM, Holt KE. 2016. In silico serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genomics* **2**.

Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6**: 90.

Jauréguy F, Carbonnelle E, Bonacorsi S, Clec'h C, Casassus P, Bingen E, Picard B, Nassif X, Lortholary O. 2007. Host and bacterial determinants of initial severity and outcome of *Escherichia coli* sepsis. *Clin Microbiol Infect* **13**: 854–862.

Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E, Lortholary O, Clermont O, Denamur E, et al. 2008. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* **9**: 560.

Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol* **52**: 1501–1510.

Johnson JR. 1991. Virulence factors in *Escherichia coli* urinary tract infection. *Clin Microbiol Rev* **4**: 80–128.

Johnson JR, Kuskowski MA, Gajewski A, Sahm DF, Karlowsky JA. 2004. Virulence characteristics and phylogenetic background of multidrug-resistant and antimicrobial-susceptible clinical isolates of *Escherichia coli* from across the United States, 2000-2001. *J Infect Dis* **190**: 1739–1744.

Kallonen T, Brodrick HJ, Harris SR, Corander J, Brown NM, Martin V, Peacock SJ, Parkhill J. 2017. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res* **27**: 1437–1449.

Kaper JB, Nataro JP, Mobley HLT. 2004. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* **2**: 123–140.

Kauffmann F. 1947. The serology of the coli group. *J Immunol Baltim Md 1950* **57**: 71–100.

Khezri A, Avershina E, Ahmad R. 2021. Plasmid identification and plasmid-mediated antimicrobial gene detection in Norwegian isolates. *Microorganisms* **9**: 52.

Kraker MEA de, Davey PG, Grundmann H, Group on behalf of the B study. 2011. Mortality and hospital stay associated with resistant *Staphylococcus aureus* and *Escherichia coli* bacteremia: estimating the burden of antibiotic resistance in Europe. *PLOS Med* **8**: e1001104.

Kudinha T, Johnson JR, Andrew SD, Kong F, Anderson P, Gilbert GL. 2013. Genotypic and phenotypic characterization of *Escherichia coli* isolates from children with urinary tract infection and from healthy carriers. *Pediatr Infect Dis J* **32**: 543–548.

Lapierre M, Blin C, Lambert A, Achaz G, Rocha EPC. 2016. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Mol Biol Evol* **33**: 1711–1725.

Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, Tenaillon O. 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. *Mol Biol Evol* **24**: 2373–2384.

Lê S, Josse J, Husson F. 2008. **FactoMineR** : An *R* package for multivariate analysis. *J Stat Softw* **25**. http://www.jstatsoft.org/v25/i01/ (Accessed July 15, 2020).

Lefort A, Panhard X, Clermont O, Woerther P-L, Branger C, Mentré F, Fantin B, Wolff M, Denamur E. 2011. Host factors and portal of entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. *J Clin Microbiol* **49**: 777–783.

Levin BR, Edén CS. 1990. Selection and evolution of virulence in bacteria: an ecumenical excursion and modest suggestion. *Parasitology* **100**: S103–S115.

Low M, Neuberger A, Hooton TM, Green MS, Raz R, Balicer RD, Almog R. 2019. Association between urinary community-acquired fluoroquinolone-resistant *Escherichia coli* and neighbourhood antibiotic consumption: a population-based case-control study. *Lancet Infect Dis* **19**: 419–428.

Manges AR, Harel J, Masson L, Edens TJ, Portt A, Reid-Smith RJ, Zhanel GG, Kropinski AM, Boerlin P. 2015. Multilocus sequence typing and virulence gene profiles associated with *Escherichia coli* from human and animal sources. *Foodborne Pathog Dis* **12**: 302–310.

Manges AR, Johnson JR, Foxman B, O'Bryan TT, Fullerton KE, Riley LW. 2001. Widespread distribution of urinary tract infections caused by a multidrug-resistant *Escherichia coli* clonal group. *N Engl J Med* **345**: 1007–1013.

Martin P, Marcq I, Magistro G, Penary M, Garcie C, Payros D, Boury M, Olier M, Nougayrède J-P, Audebert M, et al. 2013. Interplay between siderophores and colibactin genotoxin biosynthetic pathways in *Escherichia coli*. *PLOS Pathog* **9**: e1003437.

Massot M, Daubié A-S, Clermont O, Jaureguy F, Couffignal C, Dahbi G, Mora A, Blanco J, Branger C, Mentré F, et al. 2016. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of *Escherichia coli* from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology* **162**: 642–650.

Mereghetti L, Tayoro J, Watt S, Lanotte P, Loulergue J, Perrotin D, Quentin R. 2002. Genetic relationship between *Escherichia coli* strains isolated from the intestinal flora and those responsible for infectious diseases among patients hospitalized in intensive care units. *J Hosp Infect* **52**: 43–51.

Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogeneticb. *Mol Biol Evol* **30**: 1188–1195.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**: 268–274.

Nicolas-Chanoine M-H, Bertrand X, Madec J-Y. 2014. *Escherichia coli* ST131, an intriguing clonal group. *Clin Microbiol Rev* **27**: 543–574.

Nougayrède J-P, Homburg S, Taieb F, Boury M, Brzuszkiewicz E, Gottschalk G, Buchrieser C, Hacker J, Dobrindt U, Oswald E. 2006. *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. *Science* **313**: 848–851.

Nowrouzian F, Hesselmar B, Saalman R, Strannegård I-L, Åberg N, Wold AE, Adlerberth I. 2003. *Escherichia coli* in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage. *Pediatr Res* **54**: 8–14.

O'Brien CL, Gordon DMY 2011. 2011. Effect of diet and gut dynamics on the establishment and persistence of *Escherichia coli*. *Microbiology* **157**: 1375–1384.

Orskov I, Orskov F, Jann B, Jann K. 1977. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. *Bacteriol Rev* **41**: 667–710.

Östblom A, Adlerberth I, Wold AE, Nowrouzian FL. 2011. Pathogenicity island markers, virulence determinants *malX* and *usp*, and the capacity of *Escherichia coli* To persist in infants' commensal microbiotas. *Appl Environ Microbiol* **77**: 2303–2308.

Ouldali N, Varon E, Levy C, Angoulvant F, Georges S, Ploy M-C, Kempf M, Cremniter J, Cohen R, Bruhl DL, et al. 2021. Invasive pneumococcal disease incidence in children and adults in France during the pneumococcal conjugate vaccine era: an interrupted time-series analysis of data from a 17-year national prospective surveillance study. *Lancet Infect Dis* **21**: 137–147.

Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**: 3691–3693.

Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, Elion J, Denamur E. 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* **67**: 546–553.

Poolman JT, Wacker M. 2016. Extraintestinal pathogenic *Escherichia coli*, a common human pathogen: challenges for vaccine development and progress in the field. *J Infect Dis* **213**: 6–13.

Power ML, Littlefield-Wyer J, Gordon DM, Veal DA, Slade MB. 2005. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environ Microbiol* **7**: 631–640.

Qin X, Hu F, Wu S, Ye X, Zhu D, Zhang Y, Wang M. 2013. Comparison of adhesin genes and antimicrobial susceptibilities between uropathogenic and intestinal commensal *Escherichia coli* strains. *PLOS ONE* **8**: e61169.

Raimondi S, Righini L, Candeliere F, Musmeci E, Bonvicini F, Gentilomi G, Starčič Erjavec M, Amaretti A, Rossi M. 2019. Antibiotic resistance, virulence factors, phenotyping, and genotyping of *E. coli* isolated from the feces of healthy subjects. *Microorganisms* **7**.

Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* **190**: 6881–6893.

Ribera A, Roca I, Ruiz J, Gibert I, Vila J. 2003. Partial characterization of a transposon containing the tet(A) determinant in a clinical isolate of *Acinetobacter baumannii*. *J Antimicrob Chemother* **52**: 477–480.

Roer L, Tchesnokova V, Allesøe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, et al. 2017. Development of a web tool for *Escherichia coli* subtyping based on *fimH* alleles. *J Clin Microbiol* **55**: 2538–2543.

Royer G, Darty MM, Clermont O, Condamine B, Laouenan C, Decousser J-W, Vallenet D, Lefort A, de Lastours V, Denamur E, et al. 2021. Phylogroup stability contrasts with high within sequence type complex dynamics of *Escherichia coli* bloodstream infection isolates over a 12-year period. *Genome Med* **13**: 77.

Royer, G., Decousser, J. W., Branger, C., Médigue, C., Denamur, E., & Vallenet, D. 2018. PlaScope : A targeted approach to assess the plasmidome from genome assemblies at the species level. *Microbial Genomics* **4**: 9.

Russo TA, Johnson JR. 2003. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes Infect* **5**: 449–456.

Sabuncu E, David J, Bernède-Bauduin C, Pépin S, Leroy M, Boëlle P-Y, Watier L, Guillemot D. 2009. Significant reduction of antibiotic use in the community after a nationwide campaign in France, 2002–2007. *PLoS Med* **6**: e1000084.

Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. 2015. Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res* **25**: 119–128.

Schembri MA, Kjaergaard K, Sokurenko EV, Klemm P. 2001. Molecular characterization of the *Escherichia coli fimH* adhesin. *J Infect Dis* **183**: S28–S31.

Schubert S, Rakin A, Karch H, Carniel E, Heesemann J. 1998. Prevalence of the “high-pathogenicity island” of *Yersinia* species among *Escherichia coli* strains that are pathogenic to humans. *Infect Immun* **66**: 480–485.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069.

Seemann T. 2015. Snippy: rapid haploid variant calling and core SNP phylogeny. *GitHub. Available at: github. com/tseemann/snippy*.

Skurnik D, Clermont O, Guillard T, Launay A, Danilchanka O, Pons S, Diancourt L, Lebreton F, Kadlec K, Roux D, et al. 2016. Emergence of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Mol Biol Evol* **33**: 898–914.

Smati M, Clermont O, Gal FL, Schichmanoff O, Jauréguy F, Eddi A, Denamur E, Picard B, Group for the C. 2013. Real-time PCR for quantitative analysis of human commensal *Escherichia coli* populations reveals a high frequency of subdominant phylogroups. *Appl Environ Microbiol* **79**: 5005–5012.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Tedijanto C, Olesen SW, Grad YH, Lipsitch M. 2018. Estimating the proportion of bystander selection for antibiotic resistance among potentially pathogenic bacterial flora. *Proc Natl Acad Sci* **115**: E11988–E11995.

Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. *Nat Rev Microbiol* **8**: 207–217.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5**.

Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, Denamur E, Gordon D, Rocha EP. 2020. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli* ed. X. Didelot. *PLOS Genet* **16**: e1008866.

Verschuuren TD, van Hout D, Arredondo-Alonso S, Fluit AC, Reuland EA, Top J, Schürch AC, Bosch T, Bonten MJM, Kluytmans J a. JW, et al. 2021. Comparative genomics of

ESBL-producing *Escherichia coli* (ESBL-Ec) reveals a similar distribution of the 10 most prevalent ESBL-Ec clones and ESBL genes among human community faecal and extra-intestinal infection isolates in the Netherlands (2014-17). *J Antimicrob Chemother* **76**: 901–908.

Vihta K-D, Stoesser N, Llewelyn MJ, Quan TP, Davies T, Fawcett NJ, Dunn L, Jeffery K, Butler CC, Hayward G, et al. 2018. Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998–2016: a study of electronic health records. *Lancet Infect Dis* **18**: 1138–1149.

Walk ST, Alm EW, Calhoun LM, Mladonicky JM, Whittam TS. 2007. Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ Microbiol* **9**: 2274–2288.

Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG. 2004. Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonella*. *Proc Natl Acad Sci* **101**: 10644–10649.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60**: 1136–1151.

Woerther P-L, Angebault C, Jacquier H, Clermont O, Mniai AE, Moreau B, Djossou F, Peroz G, Catzeflis F, Denamur E, et al. 2013. Characterization of fecal extended-spectrum-β-lactamase-producing *Escherichia coli* in a remote community during a long time period. *Antimicrob Agents Chemother* **57**: 5060–5066.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* **67**: 2640–2644.