Lessons from twenty years of the life of commensal *Escherichia coli* populations in a human gut

Bénédicte Condamine[1], Thibaut Morel-Journel[1,2], Florian Tesson[1, 3], Mélanie Magnan[4], Aude Bernheim[3], Erick Denamur[1,5*], François Blanquart[6*], Olivier Clermont[1*]

[1] Université Paris Cité, INSERM, IAME, 75018 Paris, France
[2] Université Sorbonne Paris Nord, INSERM, IAME, 93000 Bobigny, France
[3] Institut Pasteur, Université Paris Cité, INSERM, Molecular Diversity of Microbes Lab, 75015 Paris, France.
[4] Université Paris Cité, INSERM, CNRS, Institut Cochin, UMR 1016, 75014, Paris, France.
[5] AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, 75018 Paris, France
[6] Center for Interdisciplinary Research in Biology, CNRS, Collège de France, PSL Research University, 75005 Paris, France

* ED, FB and OC shared last authorship on this work

Corresponding author: Erick Denamur, erick.denamur@inserm.fr

Running title: A 20-year period of *E. coli* in a single gut

Key words: *Escherichia coli*, commensal, trade-off, metabolism, phage, defense island, time-series

Abstract

Escherichia coli is an increasingly antibiotic-resistant opportunistic pathogen. Few data are available on its ecological and evolutionary dynamics in its primary niche, the vertebrate gut. Using Illumina and/or Nanopore technologies, we sequenced whole genomes of 210 *E. coli* natural isolates from 22 stools sampled during a 20-year period in a healthy man (ED) living in France, who did not take any antibiotics. All the phylogroups, except the C, were represented with a predominance of B2 (34.3%), followed by A and F (19% each) phylogroups. Thirty five clones were identified and classified in three phenotypes according to their abundance and persistence time: 25 sub-dominant/transient (52 isolates), five dominant/transient (48 isolates) and five dominant/resident (110 isolates). Four out of the five dominant/resident clones belonged to B2 phylogroup [sequence types (STs)131 and 452] and F-ST59, whereas sub-dominant/transient clones belonged mainly to numerous distinct B1, A and D phylogroup STs. The long residence time of B2 phylogroup clones was counterbalanced by lower colonization ability. Residence time and within-host frequency were positively correlated. By comparing ED strain genomes to a collection of commensal *E. coli* genomes from 359 French individuals, enriched/specific ED strain genome traits were identified including a metabolic pathway (*mhp* cluster) and a very rarely reported antiviral defense island encompassing dGTPase and/or Detocs systems. SNP analysis of clonal diversity showed neutral evolution. The normal *E. coli* gut microbiota is shaped by both the intrinsic properties of the strain lineages and the environmental constraints.

Introduction

Escherichia coli intestinal (diarrhea) [1] and extra-intestinal (urinary tract infection, septicemia, pneumonia, deep suppuration) [2] infections represent a major public health problem due to the frequency and the severity of these diseases. *E. coli* is also becoming increasingly antibiotic resistant [3] and has been ranked first by the WHO on the global priority list of antibiotic-resistant bacteria. However, the normal niche of *E. coli* is the gut of vertebrates, where it lives as a commensal [4] This species is thus acting mostly as an opportunistic pathogen. Selection for antimicrobial resistance happens mostly in the gut, as a commensal, and under the effect of 'bystander selection' by antibiotics [5]. Strikingly, despite the urgent need of understanding the normal ecology of *E. coli* to fight the emergence of virulence and resistance, little knowledge is available on its normal lifestyle and especially on the *E. coli* population changes over time [6], as compared to the abundant literature on *E. coli* infections. Whole genome sequences of well documented commensal isolates are dramatically lacking.

E. coli population structure and dynamics in the gut of healthy individuals has been addressed mostly in humans since the beginning of the 20th century, using in each epoch the available technologies: serotyping [7–11], multilocus enzyme electrophoresis [12,13], PCR-based approaches, pulsed-field gel electrophoresis, multilocus sequence typing, ribotyping, amplicon sequencing of the flagellin gene and shotgun metagenomics [6,14–19]. Most of these studies were performed on one to 10 subjects during a few-month period (maximum 4 years [10]) studying in general 10 colonies per sample. Some consensus emerged from these studies: (i) *E. coli* can present two extreme phenotypes associated with colonisation ability: those persisting for short periods (days to weeks), the “transients” and those persisting for long periods, the “residents” [20], (ii) a single “dominant” strain (90 to 100% within sample) is present in the majority of the samples whereas in other samples several “minor” or “sub-dominant” strains are present, (iii) the resident strains correspond to specific phylogenetic lineages often involved in extra-intestinal infections belonging to phylogroups B2 (sequence types (STs) 73, 95, 131) and F (ST59) and (iv) there is an important inter-individual variability. Similar results were obtained in non-human mammals, either in the wild in brushtail possums in Australia [21] or in cattle in Zimbabwean cows [22].

More recently, whole genome sequencing (WGS) of single isolates was used to study the evolution of a clone between the members of a household over three years [23] and we have previously studied the *E. coli* carriage of a healthy man (ED) living in Paris, France, during one year [24]. Both studies found a mutation rate per base per year about $2x10^{-7}$, in the range of what was observed in the *in vitro* long-term experiment of Richard Lenski [25]. However, at the opposite of the Lenski experiment that shows strong selection [26], *in vivo* evolution appears to be dominated by neutral changes.

In the present work, to get further insights on *E. coli* carriage over a longer time period, we extended the study of ED to 20 years (2001-2020). We analyzed 22 stool samples irregularly spaced over these 20 years, including the three first samples already published [24]. For each

new sample, we whole-genome sequenced 10 *E. coli* colonies using Illumina and/or Nanopore technologies. In addition, we compared the ED strain collection to an epidemiologically relevant collection of 359 *E. coli* isolates from the stools of 359 healthy individuals living in France and sampled between 2000 and 2017 to search for ED individual strain-specific traits.

Materials/subjects and methods

Strain samples

During a 20-year period (2001-2020), we irregularly sampled the stools of a healthy man (ED) (44-64 year-old) living in Paris, with in 2019 monthly sampling during seven months and weekly sampling during one month, nested in the monthly collect (Fig. 1 and Table S1). Ten isolates appearing as *E. coli* were randomly selected after plating the feces on Drigalski medium and stored at – 80°C in glycerol (220 isolates for 22 samples). The isolates of the three first samplings have been published previously [24]. 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).

Genome sequencing

ED isolates were sequenced using Illumina technology (short reads) on MiniSeq sequencer. All genomes were assembled using SPAdes v3.13.1 and standard parameters, and then annotated with Prokka 1.14.5 [27,28]. Sequences have been deposited at the ENA under the bioproject PRJEB45090 (Table S1). One strain from the first sample, ED1a, has also been sequenced using standard Sanger technology and assembled in two single molecules, the chromosome and the plasmid [29]. This sequence was then used in all the genomic analysis. Nanopore sequencing (long reads) was performed on one selected isolate per clone for 9 clones harboring at least 8 isolates. This sequenced isolate was selected randomly, in most of the cases within the first sample of the temporal series, and was considered as the nanopore reference isolate.

Genomes used for comparison

Commensal strains were gathered from stools of 359 healthy adults living in the Paris area or Brittany (both locations in the Northern part of France) between 2000 to 2017. One strain per individual was sequenced as above [30]. The 1,777 RefSeq genomes of high quality were downloaded from NCBI. The 113 ST399 Cplx genomes were downloaded from EnteroBase [31].

Genome analyses

Genome typing was performed as previously described [32] (https://github.com/iame-researchCenter/Petanc), including species identification, phylogrouping [33], multi-locus sequence type (MLST) determination according to the Warwick University and Pasteur Institute schemes, *in silico* serotyping [34,35] and *fimH* allele determination [36]. The presence of virulence and resistance genes was determined as

previously described with ABRicate 1.0.0, based on a custom database and Resfinder, respectively [37]. The chromosomal or plasmidic location of a gene was determined using PlaScope 1.3.1 [38]. We also searched for point mutations responsible for resistance using PointFinder 4.2.2 [39]. The phenotypic resistance pattern of the strains to main classes of antibiotics was deduced from the presence of corresponding genes or mutations using the validated approach of [40].

Pan and core genomes were computed using PPanGGOLiN 1.2.105 with default parameters [41]. The genome-wide association studies (GWAS) were performed with Scoary 1.6.16 [42] (with default parameters). A maximum likelihood phylogenetic tree was inferred based on a soft core genome alignment using IQ-TREE 2.2.0 [43] and the GTR+F+I+G4 model (1000 bootstraps) rooted on NILS31 (*Escherichia* clade I*)*. Visualization and annotation of the trees were performed with iTOL [44]. Defense system detection was done using DefenseFinder models v1.2.3 [45]. The *yjcB-ssb* defense island detection within RefSeq complete genomes (July 2022), was based on the detection of both flanking genes of more than 300 bp of the island (*pdeC* and *uvrA* genes). Alternatively, island detection in the Enterobase ST399cplx, was based on blastn 2.12.0 [46] of all the *yjcB-ssb* defense islands identified in ED and RefSeq genomes.

Illumina and nanopore contigs were assembled using Unicycler v0.4.9b [47] and annotated with Prokka 1.14.5 [28]. We then filtered the mutations identified by breseq 0.33.2 [48] as in [24]. When filtering the outputs of breseq, we looked at the predicted SNPs and short indels in the chromosomes only. We removed mutations when they were less than 51 bp apart. These clustered mutations are indeed usually caused by erroneously mapped reads.

Bayesian inference of residence times

We inferred the posterior distribution of residence time for each observed clone, then grouping clones by sequence types and by phylogroups.

We assumed the residence time of a clone is exponentially distributed with a mean residence time $T = 1/\lambda$ that we estimate. For the prior distribution of the residence time we chose a log-normal distribution with parameters 3 and 1.5. This gives a mean of approximately 20 days and a 95% higher prior density interval of 1 to 400 days, in accordance with the overall mean and range of residence times observed in the longitudinal follow-up of 8 healthy volunteers [49] .

For each episode of carriage of a clone, we defined the minimum and maximum residence times T_{max} and T_{min} from the date when it was first and last observed, and accounting for left and right censoring. The associated likelihood is the probability that the event of clearance took place in the time interval after colonization by the clone:

$$L(T) = F(T,1/T_{max}) - F(T,1/T_{min})$$

Where $F(T,\lambda)=1-\exp(-T\lambda)$ is the cumulative distribution function of the exponential distribution.

For each clone we used the associated likelihood and the prior to define the posterior distribution of T, determine the mode of the posterior distribution of carriage time, as well as the 95% higher posterior density interval. We repeated the same analysis when grouping all clones of the same ST together assuming independence between carriage times of these distinct clones. We repeated the same analysis when grouping all clones of the same phylogroup together.

Many clones are observed only once. These clones have T_{min}=0 but do not have the same T_{max}. However, in all cases the posterior mode of residence time was inferred to be at $\check{T}$=2.1 days.

Analysis of the within-clone diversity

Within-clone diversity was assessed by computing Π_{ij} as the average number of mutations per site separating two isolates of the same clone, noted *i* and *j* [50]. Hence, the average number of mutations separating two isolates of the same clone and of the same sample *i* was noted Π_{ii}. The number of mutations accumulated per site over *T* days separating samples *i* and *j* with a mutation rate *v* could then be estimated as $\nu T = \Pi_{ij} - \Pi_{ii}$ [50]. νT was computed for the 12 samples that were not the first for a clone (58 isolates), and a linear regression over time was performed to estimate the mutation rate *v* per site per day.

Under neutrality, the divergence between two clones depends only on the mutation rate per site per generation (μ) and the effective population size (Ne), such that $\Pi_{ii} = \mu Ne$ when eventually reaching the mutation-derive equilibrium. We simulated a Wright-Fisher model over 1065 days, with identical genotypes (Π_{ii}= 0) at t = 0 single nucleotide mutations occurring along the chromosome with a probability μ [50].

Results

Twenty-two stool samples from the ED subject, corresponding to 220 isolates (10 isolates per sample), distributed on a 6,949 day-period (24 October 2001 to 2 November 2020) were analyzed (Fig. 1). The intervals between the samples ranged from seven days to 1,966 days/5.38 years (mean 331 days). Importantly, the ED subject did not take any antibiotics nor had been hospitalized during the sampling period. Twenty one of the 30 sequences from the first three samples [24] were available (Table S1). For the nine missing sequences, the corresponding isolates were sequenced together with the 190 remaining isolates, in a first step by short read sequencing (Illumina technology). Except for one sample from 2011 which *f*yielded 10 isolates of a unique clone of *Hafnia alvei*, all other samples were positive for *E. coli stricto sensu* [51], leading to a collection of 210 *E. coli* isolates (Table S1).

Phylogenomic analysis of the ED *E. coli* isolates and delineation of clones

Clonal diversity within the 210 ED E. coli *isolates*

In silico phylotyping, confirmed by a SNP based core genome phylogeny (Fig. S1), showed that all the known phylogroups [51], except the C phylogroup, were represented among the 210 isolates (Table 1). The most prevalent was the B2 phylogroup (N=72 isolates, 34.3%), followed by the A and F phylogroups (N=40, 19% each), the B1 phylogroup (N=26, 12.4%), the D phylogroup (N=17, 8%) and the H, G and E phylogroups (5, 1.9 and 0.5%, respectively). The most frequent STs were the STs 131 (20%, B2 phylogroup) and 59 (19.5%, F phylogroup) followed by the ST10 (11%, A phylogroup) (Table S2).

Table 1. Distribution of the 210 *E. coli* isolates gathered from the stools of the individual ED

Phylogroup	Nb (%) of isolates	Nb (%) of clones	Isolate/clone ratio
A	40 (19)	11 (31.4)	3.6
B1	26 (12.4)	10 (28.5)	2.6
B2	72 (34.3)	4 (11.4)	18
D	17 (8)	6 (17.1)	2.8
E	1 (0.5)	1 (2.8)	/
F	40 (19)	1 (2.8)	40
G	4 (1.9)	1 (2.8)	/
H	10 (5)	1 (2.8)	/

The clones were first defined as distinct haplogroups (combination of the phylogroup, the ST with both Warwick University and Institut Pasteur schemes, the serotype O:H, the *fimH* allele) and then confronted to the number of SNPs between isolates within the core genome [52]. Thirty four haplogroups were identified with a number of inter-haplogroup isolate SNPs ≥ 130 (Table S3). The number of SNPs intra-haplogroup was smaller or equal to 51, except in the phylogroup A_ ST399/661_O25:H12_fimHND haplogroup, where isolates differed by 125 to 130 SNPs. We considered that this haplogroup encompassed two different clones, named A and B. With our approach, we were thus able to identify 35 clones named thereafter "ED clones" among the 210 *E. coli* isolates, considering all the ST452 isolates as a single clone [24] (Table S4, Table 1). The main STs among the 35 clones were ST10 (11.4%) followed by ST131 and 69 (D phylogroup) (8.6% each).

The distribution of these clones in the phylogroups was very uneven. Excluding the rarely identified E, G and H phylogroup isolates, two categories of clones can be distinguished. On one hand, numerous clones are identified among the isolates (isolate/clone ratio from 2.8 to

3.6) in phylogroups B1, D and A, indicating a high genetic diversity whereas very few clones (isolate/clone ratio from 18 to 40) were identified in phylogroups B2 and F, respectively (Table 1).

ED isolate core and pan genomes

The soft core genome of the 210 isolates (*i.e.*, genes present in 95% to 100% of the isolates) was composed of 3,278 genes. We took this definition of core genome as it provides some flexibility to account for sequencing or assembling artifacts. On the other hand, the pangenome (the totality of genes) encompassed 13,627 genes. The soft core and pangenome of the clones were similar to that of the isolates, although slightly smaller (3,270 and 12,986, respectively).

Global genomic characteristics and virulence, antibiotic resistance and antiviral defense island gene content of the isolates overtime

We first asked whether the global genomic characteristics of the isolates were constant in time. The mean of the total number of genes per isolate at a given day and the mean of the percentage of plasmid genome length among the complete genome were stable over years with a global mean at 4,768.27 and 2.30 %, respectively. The number of replicons per isolates was also relatively stable (global mean 3.83), with a clear predominance of Col and IncF plasmids (Fig. S2).

We then focused on specific genetic elements known to be important in the survival of *E. coli* in its commensal lifestyle in the gut: the virulence associated genes, as it has been proposed that virulence is a by-product of commensalism [53], the antibiotic resistance genes, as antibiotics are major selective agents and the defense islands against phage predation [54]. These factors may be primarily colonization factors [52]. We also compared the ED clones to a control collection of 359 *E. coli* isolates from the stools of 359 healthy individuals living in France during the same period, to potentially identify ED specific features.

When considering functional classes of virulence factors, i.e. adhesins, iron acquisition systems, protectins, toxins and invasins, we did not see any significant difference over time (Fig. 2A). Using random samples of 35 strains from the control commensal collection, we found the ED clones had significantly less virulence genes than the commensals (mean 47 in ED against 56 in commensals, $p = 0.00014$ according to a permutation test).

Genes or mutations conferring antibiotic resistance appeared briefly in 2010 in a ST362 strain exhibiting bla_{TEM1B}, *sul2*, *tet(A)* and *catA* genes and were more prevalent starting in 2018, with five clones bearing resistance determinants: ST131 O16 (bla_{TEM1B}, *aac(3)-IId*, GyrA 83), ST10 O89 (bla_{TEM1B}, *sul1, sul2, dfrA5, tet(A)*), ST69 O23 (bla_{TEM1B}), ST1086 (*fosA*), ST399 O129 (*tet(B)*), conferring resistance to ampicillin, aminoglycosides, quinolones, sulfonamides, fosfomycin, tetracyclines and chloramphenicol (Fig. 2B). No ESBL coding genes were detected. As the subject did not use antibiotics during the study period, these

resistant strains were able to colonize in the absence of antibiotic selection. Using random samples of 35 strains from the control commensal collection, we found the number of resistance genes in the ED clones was slightly less than typical in commensals (mean 0.6 in ED against 1.1 in commensals, p = 0.045 according to a permutation test) (Fig. S3).
Finally, we searched exhaustively for the known defense systems in the genomes of the isolates using DefenseFinder [45]. In all the 35 ED clones, we found 74 different defense system types (7.8 mean per bacterium +\- 2.5 [3-13]). We did not see any trend over time for the number of defense systems, nor the diversity of defense (Fig. 2C). Using random samples of 35 strains from the control commensal collection, we did not find any significant difference in the diversity (number of different systems types) of defense systems with the ED clones (74 in total in ED against 69 in commensals, p = 0.87 according to a permutation test) (Fig. S3). However, two systems were significantly more abundant in the ED clones than in the control commensal strains (chi2 test corrected by Bonferroni): dGTPase and Detocs (Fig. S4). We will come back to this finding below in a more systematic analysis.

Epidemiological dynamics of ED clones

We then examined the epidemiological dynamics of the ED clones which depend on how long clones persist in ED, and on the rate of colonization of the strains able to colonize ED.

Persistence times and dominance of ED clones

We quantified the residence time of ED clones with a Bayesian approach incorporating prior information on residence times measured in a previous study [49]. We inferred the posterior distribution of the residence time for each observed clone (Table S5). Most of the clones (30 out of 35) were seen only once and consequently had an inferred posterior residence time of 2.1 days, with the remaining five clones classified as resident (residence time superior to 2.1 days, with posterior mode ranging from 10.1 to 437.1 days). Moreover, we classified clones by their dominance. Clones were considered dominant if present at least in one sample with at least 5 isolates; otherwise, they were classified as sub-dominant. Most of the clones (25 over 35) were sub-dominant with a mean of 2.1 isolates per sample.

All in all, among the 35 clones, 25 were sub-dominant/transient (52 isolates), five were dominant/transient (48 isolates) and five were dominant/resident (101 isolates) (Table 2). Residence time was positively correlated with within-host frequency (linear model, effect size = 93 days, 95% CI [2.1-180], p = 0.045) (Fig. 3A). Phylogroups and STs were unevenly distributed among categories with B2 (STs 131 and 452) and F (ST59) phylogroups found in most of the dominant/resident strains but never as subdominant/transient. Conversely, phylogroup B1 (represented by numerous distinct STs), and to a lesser extent A and D, were found in subdominant/transient strains. Among the 21 samples, a dominant clone was found in 15 cases (71.4%) with in nine cases a full dominance (10 over 10 isolates) whereas in the six remaining samples, three to six clones were present per sample. The minimum duration of carriage of resident clones varied from seven days (ST635 O11) to 28 days (ST131 O25 B),

315 days (ST452), 489 days (ST59) and 1065 days (ST131 O25 C1), corresponding to inferred posterior modes of 10, 163, 234, 437 days respectively (Fig. 1).

Table 2. Feces carriage categories of the 210 isolates belonging to the 35 identified clones

Strain category*	Nb (%) of isolates	Nb (%) of clones	Isolate /clone ratio	Clone phylogroup repartition** Nb (%)	Global virulence score (mean)	Adhesin score (mean)	Iron capture system score (mean)	Invasin score (mean)	Toxin score (mean)	Miscellaneous score (mean)	Resistance score (mean)
Subdominant / Transient	52 (25.9)	25 (71.4)	2.1	A: 8 (72.7) B1: 10 (100) G: 1 (100) E: 1 (100) D: 5 (83.3)	42.0	15.3	16.2	0.3	6.2	1.6	1.1
Dominant / Transient	48 (23.9)	5 (14.3)	9.6	A: 2 (18.2) H: 1 (100) B2: 1 (25) D: 1 (16.7)	59.2	23.1	22.7	1.0	6.9	1.8	2.3
Dominant / Resident	110 (50.2)	5 (14.3)	22.0	F: 1 (100) B2: 3 (75) A: 1 (9)	60.4	22.1	25.0	0.9	5.2	2.5	0.5

* See the text for definitions, ** The percentages are calculated per phylogroup.

Association between persistent time and virulence gene content

Long term intestinal persistence of colonizing *E. coli* has been associated with the presence of specific clones belonging mainly to B2/D/F phylogroups, these clones exhibiting numerous virulence associated genes, especially those involved in iron capture systems [17,55]. Fluoroquinolone resistance has also been reported to be associated with long-term carriage [56]. We thus investigated the association between the three categories of clones (sub-dominant/transient, dominant/transient, dominant/resident) and the score of virulence and resistance associated genes (Table 2).

The dominant/resident clones had the highest global virulence score while the sub-dominant/transient clones had the lowest virulence score (ANOVA, p=0.055). This is particularly true for the adhesin and iron capture system functional classes. When considering the predictors of residence time, only the within-host frequency but not the number of adhesins, iron capture systems, and toxins, significantly positively correlated (Fig. 3 A,B). The dominant/transient clones have the highest resistance score (Kruskal-Wallis, p=0.0028).

Colonization rates of ED clones and the trade-off with persistence

The epidemiological dynamics of *E. coli* rely not only on the intestinal persistence of clones, but also on their ability to colonize new intestines, which was assessed at the phylogroup level. The number of colonization events (i.e. the first appearance of a new *E. coli* clone) was computed for phylogroups represented by more than a single clone (A, B1, B2, D) and compared with the expected number of events assuming an equal colonization rate for all phylogroups, accounting for the relative phylogroup frequencies in France as estimated from the 359 *E. coli* control isolates from healthy individuals [57]. Colonizations of the gut of ED by B1 and B2 phylogroups were respectively more frequent (p = 0.0301) and rarer (p = 0.0034) than expected, suggesting intrinsic differences in colonization rates (Fig. 3C). Interestingly, the high residence time of B2 clones is counterbalanced by a lower colonization ability.

Genomic features associated with clones successfully colonizing ED

To identify further potential distinct genetic features of *E. coli* strains able to colonize the focal individual, we compared the core and pan genomes of *E. coli* harbored over 20 years by the individual to the epidemiologically relevant control collection of 359 *E. coli* strains [57]. ED specific/enriched genes may reflect genes allowing adaptation to the genetics or lifestyle of the focal individual.

Identification of ED-specific genes

The soft core genome of the control collection was composed of 3,262 genes whereas the pan genome encompassed 27,752 genes. We compared the soft cores of the 35 ED clones and the

commensal control collection. We found 3,198 genes in common and identified 72 and 64 genes specific to the ED and control collections, respectively. Among the 72 genes found only in the soft core of ED clones, we identified a cluster of eight genes, *mhpA* to *F*, *mhpT,* coding for enzymatic functions allowing the use of the 3-(3-hydroxyphenyl)propionate (3HPP), one of the phenylpropanoids from lignin [58] (see below). Three additional genes were identified: *ldrA*, coding for a small toxic polypeptide [59], enriched in motile strains [60], *rtcB*, coding for a RNA-splicing ligase that afford bacteria a means to recover from stress-induced RNA damage [61] and a type III secretion system effector (EspL1) coding gene (Table S6).

Systematic genome wide association study (GWAS) of ED versus *control commensal strains*

To go further, we performed a systematic GWAS with the software Scoary to investigate if some genes were more generally differentially retrieved between the 35 ED and 359 control commensal pangenomes. The software scores all genes of the pan-genome for associations while accounting for population stratification [42]. We found 36 genes significantly associated with the 35 ED clones (Table 3). In depth analysis of some biologically relevant genes is presented below.

Table 3. List of genes significantly associated to the 35 ED clones

Gene*	Annotation	Presence in 35 ED clones n (%)	Presence in 359 commensals n (%)	Naive p-value	Corrected p-value
tufB	Elongation factor Tu 1	21 (60%)	73 (20.3%)	1,61E-06	1,41E-02
mazE	Antitoxin MazE	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
ogg	N-glycosylase/DNA lyase	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
queC	7-cyano-7-deazaguanine synthase	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
RBKS	Ribokinase	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
group_21776	hypothetical protein	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
group_21778	hypothetical protein	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
group_39802	hypothetical protein	5 (14.3%)	0 (0%)	4,21E-06	1,41E-02
group_39314	hypothetical protein	7 (20%)	5 (1.4%)	1,38E-05	3,49E-02
group_11850	hypothetical protein	13 (37.1%)	31 (8.6%)	1,81E-05	3,49E-02

mhpC	2-hydroxy-6-oxononadienedioate/2-hydroxy-6-oxononatrienedioate hydrolase	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
mhpD	2-keto-4-pentenoate hydratase	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
mhpE	4-hydroxy-2-oxovalerate aldolase	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
mhpB	2,3-dihydroxyphenylpropionate/2,3-dihydroxicinnamic acid 1,2-dioxygenase	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
mhpT	3-(3-hydroxy-phenyl)propionate transporter	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
mhpF	Acetaldehyde dehydrogenase	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
mhpA	3-(3-hydroxy-phenyl)propionate/3-hydroxycinnamic acid hydroxylase	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
group_10295	hypothetical protein	35 (100%)	257 (71.6%)	2,35E-05	3,49E-02
rhsA	hypothetical protein	9 (25.7%)	13 (3.6%)	2,68E-05	3,49E-02
group_17650	hypothetical protein	7 (20%)	6 (1.7%)	2,81E-05	3,49E-02
dcuD	Putative cryptic C4-dicarboxylate transporter DcuD	7 (20%)	6 (1.7%)	2,81E-05	3,49E-02
kpsM	Polysialic acid transport protein KpsM	7 (20%)	204 (56.8%)	2,86E-05	3,49E-02
rsxC	Ion-translocating oxidoreductase complex subunit C	17 (48.6%)	293 (81.6%)	3,38E-05	3,94E-02
metC	Cystathionine beta-lyase	6 (17.1%)	4 (1.1%)	5,23E-05	4,19E-02
group_21241	hypothetical protein	6 (17.1%)	4 (1.1%)	5,23E-05	4,19E-02
ygeK	hypothetical protein	6 (17.1%)	4 (1.1%)	5,23E-05	4,19E-02
group_40459	hypothetical protein	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
group_11090	hypothetical protein	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
group_21782	hypothetical protein	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
group_21388	hypothetical protein	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
group_39799	DtcB	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
*group_*	DtcC	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02

39800					
group_ 39798	DtcA	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
group_ 39796	hypothetical protein	4 (11.4%)	0 (0%)	5,30E-05	4,19E-02
group_ 18898	hypothetical protein	1 (2.9%)	116 (32.3%)	6,01E-05	4,61E-02
yjgL	putative protein YjgL	12 (34.3%)	30 (8.4%)	6,37E-05	4,76E-02

* The genes in green and blue belong to the *yjcB-ssb* defense island and the *mhp* operon, respectively, whereas the *kps* gene is in orange.

First, the *kpsM* gene involved in the biosynthesis of the K1 capsule was under-represented in the ED clones. Second, one significant cluster of genes was the *mhp* gene cluster already identified above as an ED-specific soft genome gene (Table 3 - highlighted in blue). It was found in all 35 ED clones but only in 71.6% of the control strains. An association with specific STs has been recently reported for this gene cluster [62], present in ST131, but not in other main B2 phylogroup STs. It is in fact also present in the B2 phylogroup ST452, the first ED clone, rendering all the ED B2 strains positive for this cluster and explaining its subsequent identification by GWAS. Third, a cluster composed of 15 genes corresponding to an antiviral defense island [45], was found to be present in five clones (in complete and incomplete form in four and one clones, respectively) but never in the 359 control strains (Fig. 4A, Table 3 - genes highlighted in green). This island (18 kbp) was inserted between *pdeC-yjcB* and *uvrA-ssb* core-genes and thereafter named the *yjcB-ssb* defense island. It corresponds to the integration hotspot 39 [63]. At least two defense systems were present inside the complete form: (i) a Detocs system composed of three genes, coding for a two-component system composed of an histidine kinase containing protein (DtcA), a “buffer” protein that absorbs phosphate signals (DtcB) and a response regulator (DtcC) [64]. (ii) a *dGTPase* gene that encodes deoxyguanosine triphosphatase (dGTPase) enzyme, which degrades dGTP into phosphate-free deoxyguanosine, depleting the nucleotide pool and halting the phage replication [65]. This gene is adjacent to (iii) *mazE*, coding for the antitoxin of the MazF/MazE module that suppresses phage lytic propagation via abortive infection [66]. In the incomplete form of the island, Detocs genes are replaced by DNA mobility genes. In addition, one isolate/clone (ED2020-11-02-i) encodes for a different Detocs system (DtcC 48.1% similarity, DtcB 58.5% similarity, DtcA 37% similarity with Detocs system from ED2020-11-02-f), but in a different region of the genome.

To characterize the phylogenetic context of the six clones carrying the defense island and/or Detocs system, we mapped these strains on a core genome tree reconstructed using the 35 ED clones and the control commensal strains. The *yjcB-ssb* defense island positive strains belonged to two closely related lineages (ST635/2531 and ST399), also called ST399 Cplx, within the phylogroup A0 (*arpA* +, *chuA* -, *yjaA* -, TspE4.C2 -) [67]. They are intertwined with the ED2020-11-02-i isolate and with island-negative control strains (Fig. 4A). When

searching for this island in 1,777 *E. coli* genomes of high quality from the RefSeq database, we found only two positive strains for the complete island belonging to phylogroup/ST A ST635 and B1 ST58 (Fig. 4B). SNP comparison of 12kb of the defense island (present in a single contig in all clones) between the different isolates showed very few SNPs among the ED clones (maximum 2 SNPs) (Fig. 4B), but a little more between the islands in the six ED clones and the island in the RefSeq A ST635 strain (between 7-8 SNPs), the B1 strain island being more phylogenetically distant (between 9 and 11 SNPs). As the defense island seemed to be restricted mainly to the ST399 Cplx, we searched for its presence in 113 *E. coli* ST399 Cplx genomes from EnteroBase [31]. Indeed, we found that half of the strains (51.3%), scattered in the core genome phylogenetic tree, possessed the *yjcB-ssb* island either in its complete (24.8%) or incomplete (26.5%) form (Fig. S5), confirming the link between this island and the ST399 Cplx.

The *yjcB-ssb* defense island was observed for the first time in September 2019 (day 6522) and found subsequently in November and December 2019 (days 6608 and 6615), and June and finally November 2020 (days 6803 and 6949) (Fig. 1). If we consider individually the two main defense systems, the ST635 O11 and ST635 O17 clones are only dGTPase and Detocs positive, respectively.

Intra-host evolution of resident ED clones: absence of sign of selection and mutation rate

In previous work studying the three first samples of the time series, we analyzed the genomic diversity of the ST452 clone at the SNP level using Illumina technology coupled to a circularized genome of high quality (ED1a) [24,29]. We pursued this analysis on nine additional clones (128 isolates) (Table S7) corresponding to unique samples in five cases where the clone was dominant (8 to 10 isolates per sample) and to four clones (8 to 40 isolates per clone) sampled at multiple dates, with seven to 1065 days separating the first and last samples. We used Illumina technology (short reads) on all isolates associated with one isolate sequenced with Nanopore technology (long reads) per clone. Using breseq 0.33.2 [48] we identified 48 different mutations, 43 of them SNP and 33 of them genic SNP. No parallelism was observed at the gene level, as all the mutations targeted distinct genes (Table S7). The ratio of non-synonymous-to-synonymous mutations was estimated, based on 20 codon usage frequency tables established for *E. coli* from international DNA sequence databases [56,68] including more than 1,000 codons. The ratios computed varied between 1.125 and 1.370, consistent with neutral evolution.

Of note, the clone F ST59/815 O1:H7 *fimH*34 presented two distinct haplotypes, both already present in the first sample and differing from each other by six mutations. The two haplotypes were found in later isolates (the first one 92, 134, 141 and 155 days later; the second one 30 and 489 days later). To account for this, these two haplotypes have been considered separately in the following analysis of substitution rates. A clustering analysis was used to assign each isolate to one of the two haplotypes, with 67.9% of the total variance observed for this clone being accounted for by the separation in two haplotypes.

The mutation rate of *E. coli* in the ED samples was then estimated as the slope of a regression of the number of new mutations accumulated by each clone over time [69], at 1.503×10^{-9} mutation per day per nucleotide (95%CI = [1.159×10^{-9}, 1.848×10^{-9}]), i.e. 5.491×10^{-7} mutation per year per nucleotide or 2.77 mutations per year per genome (Fig. 5A). Based on previous estimations of the mutation rate at 2.2×10^{-10} mutation per generation per nucleotide [69], the estimated generation time of *E. coli* in the ED samples is therefore 3.51 hours (95%CI = [2.86, 4.56]). Given this generation time, a single ingested *E. coli* could reach densities of the order of 10^8 cells/g of feces in around five days (128.3 h), assuming exponential growth and a colon holding 1 kg of feces. Assuming neutrality, only the mutation rate and the effective size (Ne) of the *E. coli* population should define the divergence between two isolates of the same clone over time. Using a stochastic model describing the appearance and fixation of mutations, we simulated the number of mutations separating two initially identical clones over time, for the estimated mutation rate and various effective population sizes. The divergence observed in ED is consistent with Ne = 1000, the model predicting an equilibrium reached in about a year (Fig 5B).

Discussion

This work is based on the originality of our collections with (i) the strains from the ED individual that can be considered as "wild" commensals of a high income country due to the absence of medical records of the host, as confirmed by the low level of virulence and resistance genes and (ii) the commensal strains from the French subjects that were used as a control cohort. The most striking features are discussed below.

Large diversity of *E. coli* hosted by a single individual

Within a single individual, we observed a large diversity of STs, and all phylogroups were represented. As a consequence, the soft core genome of *E. coli* in a single individual (n=3,270), was almost identical to the control commensal collection of trains from 359 individuals (n=3,262). When comparing the ED core genome to a more diverse datasets chosen to reflect the diversity of the whole species, represented by 1,294 Australian strains from various origins (mammal and bird hosts, environment) selected from more than 5,000 strains, we also found near identical core genomes (n=3,242) [70] (data not shown). The pangenome was composed of more than 13,000 genes. Considering that the human genome encompasses 20,000 protein-coding genes [71], this means that the individual hosted a number of *E. coli* genes representing almost 70% of the number of his own genes.

The gut colonization/persistence phenotypes are phylogroup/lineage dependent

We were able to demonstrate two paces of strain turnover in the human gut: rapid turnover with a timescale of the order of one week, for highly diverse B1 and most A clones (sub-dominant and transient phenotypes); and slow turnover with a timescale of the order of one year, for a few extra-intestinal pathogenic *E. coli* belonging to B2 (ST131) and F (ST59)

phylogroups (dominant and resident phenotypes). The longest observed period of carriage was three years (ST131 O25 C1). We pursued the ED stool analyses in an ongoing research project and found the ST59 strain as dominant in the stools of ED in December 2021, corresponding to 2 years and 5 months (884 days) of asymptomatic carriage (Blanquart and Clermont, personal data).

The diversity in a single subject in the present study was in agreement with what was previously found in a large cohort of 100 healthy French subjects with 100x higher threshold of detection of the four main phylogroups [72]. The authors showed that, when phylogroup B2 was dominant, it tended not to co-occur with other phylogroups. In contrast, other phylogroups were present when phylogroup A was dominant. The dominant B2 clones belonged mainly to archetypal ExPEC ST95, 131 and 73 [51].

The link between B2/D phylogroups and long residence time in the gut has been reported in analyses of a longitudinal study of 70 Swedish infants based on the identification of the four main phylogroups [73]. Using the more complete scheme of the Clermont typing method, Martinson *et al*. showed in a smaller and more detailed longitudinal cohort study (8 adult participants, 8 to 30 months) that A (corresponding mainly to ST10 complex), B2.3 (corresponding to the majority of B2 ExPEC) and F phylogroups are more likely to be human residents compared to the other genotypes [49]. Similarly, longitudinal cohort study of neonatal gut microbiomes showed that ExPEC clones (ST73, 131, 69, 95 and 59) were the best gut colonizers within the first 21 days, tended to dominate the *E. coli* flora and could reside up to one year [74]. A re-analysis of the data from Martinson and Walk [49] showed that the long persistence of some phylogroups was counterbalanced by a poorer capacity to colonize hosts. This persistence-colonization trade-off, together with an additional mechanism of niche differentiation, could in theory explain the coexistence of multiple *E. coli* phylogroups in the population [75].

As there is a link between B2 strains and the abundance of virulence genes [76], it is difficult to untangle the respective roles of the phylogenetic background and virulence genes. A recent study on 222 American veterans and household members [17] showed that the ST131-O25-*fimH30* was the strongest univariate correlate of intestinal persistence. However, in the multivariate analysis, multiple virulence genes (especially the iron uptake systems) and fluoroquinolones were better predictors of persistence times than the ST131 background. Our data, possibly because of the small sample size, were also non conclusive. Larger datasets on the duration of persistence in the gut, together with whole genome sequences, would enable a GWAS that may be powered to untangle effects of individual variants from the genetic background.

A key unanswered question is how such diverse B1 and at a lesser extent A phylogroup strains can colonize the ED stools (often two different clones per sample) knowing that carriage of B1 strains in humans in France is rare (13%) [77], B1 strains being commonly found in cattle [78] and also in water and sediments [79].

Host specific traits in the *E. coli* genomes: role of metabolism and phage predation

Beside the genomic diversity of the ED clones, we identified genes significantly enriched in the individual as compared to a pool of 359 individuals living in similar conditions. Interestingly, these genes are involved in two fundamental processes of gut colonization: metabolism [80] and anti phage defense [54].

The gene cluster encoding the 3-(3-hydroxyphenyl)propionate (3-HPP) catabolic pathway is enriched in ED strains, particularly in resident B2 strains from ST452 and ST131. This pathway allows strains to use aromatic acids such as phenylpropanoids as sole sources of carbon for growth. Such compounds are the result of the action of intestinal microflora on cereal constituents [81]. It is tempting to speculate that the enrichment in *mhp* genes in ED is linked to his high and daily consumption of cereals. Further microbiota studies should be performed in relation to diet to test the hypothesis.

Bacteria protect themselves from foreign DNA as phage DNA by more than a hundred systems, unveiling many different molecular mechanisms 54[54]: restriction systems [82], CRISPR-Cas systems [83], although their roles in natural *E. coli* isolates remain elusive [84], and dozens of novel systems recently identified by high-throughput screening methods [85]. A majority of these newly identified systems have been uncovered through the "defense island" method, using a guilt by association approach, without epidemiological and/or experimental validation. We showed here that six clones belonging to the ST399 Cplx and bearing the very rarely reported dGTPase and/or Detocs defense systems are present during the last year of the study, arguing for their selection in the presence of a phage sensitive to the Detocs and/or dGTPase system(s) within the fecal phageome of ED individual. Indeed, although closely related, these clones came from distinct colonization events as attested by the number of SNPs between them (140 to 10,000, Table S3). However, no evidence on which phage(s) are involved in this selection is available in the literature [86]. These clones are sub-dominant except one (Fig. 1), in agreement with data in the mouse gut showing that phages knockdown and coexist with targeted bacteria, leading to mixed populations of resistant and susceptible bacteria with enrichment of a phage-resistant subpopulation [87]. The presence of this rare defense island in a specific lineage is striking and could result from epistatic interactions at the genome level as recently reported in iron capture systems [88]. Moreover, the presence of this rare island in six distinct successful colonization events suggests strong selection acting at the between-host level. The island is present in none of the 359 control strains, and in two out of 1,777 high-quality *E. coli* genomes. The frequency of the island in the natural *E. coli* population could thus be of the order of 0.1%. The occurrence of six events of colonization with strains carrying this island implies that at least ~6000 *E. coli* strains contended to colonize ED over the 14 months when the six strains successfully colonized (Sep 2019 - Nov 2020), resulting in a rate of at least ~14 *E. coli* strains per day contending to colonize a host. These observations suggest that "Everything is everywhere, but the environment selects" [89]. Individual hosts constantly ingest large numbers of contending *E. coli* strains. Only a small fraction of them successfully colonize. Successful

colonization can result from the adaptation of specific genotypes to the particular environmental conditions in the gut at this time.

No sign of molecular adaptation over time within the gut of a single individual

It has been reported that *E. coli* isolates are adapting under strong selective pressure when colonizing an extra-intestinal site. This was based on genomic signs of selection, such as a high nonsynonymous to synonymous mutation ratio and mutational convergence [90]. At the opposite, no molecular footprint of selection has been identified in commensal *E. coli* isolates from healthy individuals [91], including the previously published ST452 clone isolated from ED [24]. The additional data that we obtained from ED isolates confirm this apparent absence of selection at the within-host level. Similar to the previous inferences on the ST452 clone [24], we found a mutation rate of 2.77 mutations per year per genome with an estimated generation time of *E. coli* of 3.51 hours in the new ED samples.

Interestingly, two haplotypes of clone F ST59/815 O1:H7 *fimH*34 identified were present together in the first sample of this clone, but not in the subsequent ones. Afterwards, samples contained only one or other haplogroup, but neither haplotype was seemingly able to eliminate the other. Indeed, one was detected up to the penultimate sampling (155 days later), and the other at the last (489 days later). This suggests coexistence between the two closely related haplotypes, with competition generating opposite fluctuations in bacterial densities over time, so that only the most abundant one at a given sampling date is detected. There is experimental evidence for coexistence between very-closely related but competing *E. coli* strains because of negative frequency-dependent effects [92], which were previously identified among *E. coli* [93] and are consistent with the fluctuations observed.

Conclusions

Despite originating from a single individual, our data strongly suggest that the normal *E. coli* microbiota is shaped by both the intrinsic properties of clonal lineages and environmental constraints such as nutrient availability and presence of phages. Such long term time series studies of *E. coli* microbiomes in a large range of subjects worldwide should be supported to see how our data are generalisable.

Acknowledgments

We are grateful to Julie Marin and Marie Petitjean for providing the control commensal and EnteroBase data sets, respectively. We thank Guillaume Achaz for pointing out the method to analyze SNP diversity in longitudinal samples.

References

1 Khalil IA, Troeger C, Blacker BF, *et al.* Morbidity and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990-2016. *Lancet Infect Dis*. 2018;18:1229–40.
2 Feldman SF, Temkin E, Wullfhart L, *et al.* A nationwide population-based study of Escherichia coli bloodstream infections: incidence, antimicrobial resistance and mortality. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. 2022;28:879.e1-879.e7.
3 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet Lond Engl*. 2022;399:629–55.
4 Tenaillon O, Skurnik D, Picard B, *et al.* The population genetics of commensal Escherichia coli. *Nat Rev Microbiol*. 2010;8:207–17.
5 Tedijanto C, Olesen SW, Grad YH, *et al.* Estimating the proportion of bystander selection for antibiotic resistance among potentially pathogenic bacterial flora. *Proc Natl Acad Sci U S A*. 2018;115:E11988–95.
6 Martinson JNV, Walk ST. Escherichia coli Residency in the Gut of Healthy Human Adults. *EcoSal Plus*. 2020;9. doi: 10.1128/ecosalplus.ESP-0003-2020
7 Totsuka K. Studien über Bacterium coli. *Z Für Hyg Infekt*. 1903;45:115–24.
8 Wallick H, Stuart CA. Antigenic Relationships of Escherichia coli Isolated from One Individual. *J Bacteriol*. 1943;45:121–6.
9 Perch B. Weitere Untersuchungen über Die Coliflora Des Gesunden Menschen. *Acta Pathol Microbiol Scand*. 1944;21:239–47.
10 Sears HJ, Brownlee I. Further observations on the persistence of individual strains of Escherichia coli in the intestinal tract of man. *J Bacteriol*. 1952;63:47–57.
11 Robinet HG. Relationship of host antibody to fluctuations of Escherichia coli serotypes in the human intestine. *J Bacteriol*. 1962;84:896–901.
12 Caugant DA, Levin BR, Selander RK. Genetic diversity and temporal variation in the E. coli population of a human host. *Genetics*. 1981;98:467–90.
13 Tzabar Y, Pennington TH. The population structure and transmission of Escherichia coli in an isolated human community; studies on an Antarctic base. *Epidemiol Infect*. 1991;107:537–42.
14 McBurney W, Mccartney A, Apun K, *et al.* Perturbation of the enterobacterial microflora detected by molecular analysis. *Microb Ecol Health Dis*. 1999;11. doi: 10.3402/mehd.v11i3.7902
15 Johnson JR, Clabots C, Kuskowski MA. Multiple-host sharing, long-term persistence, and virulence of Escherichia coli clones from human and animal household members. *J Clin Microbiol*. 2008;46:4078–82.
16 Damborg P, Nielsen SS, Guardabassi L. Escherichia coli shedding patterns in humans and dogs: insights into within-household transmission of phylotypes associated with urinary tract infections. *Epidemiol Infect*. 2009;137:1457–64.
17 Johnson JR, Clabots C, Porter SB, *et al.* Intestinal Persistence of Colonizing Escherichia coli Strains, Especially ST131-H30, in Relation to Bacterial and Host Factors. *J Infect Dis*. 2022;225:2197–207.
18 Hu D, Fuller NR, Caterson ID, *et al.* Single-gene long-read sequencing illuminates Escherichia coli strain dynamics in the human intestinal microbiome. *Cell Rep*. 2022;38:110239.
19 Han N, Peng X, Zhang T, *et al.* Rapid turnover and short-term blooms of Escherichia coli in the human gut. *J Bacteriol*. 2024;206:e0023923.
20 Sears HJ, Brownlee I, Uchiyama JK. Persistence of individual strains of Escherichia coli in the intestinal tract of man. *J Bacteriol*. 1950;59:293–301.
21 Blyton MDJ, Banks SC, Peakall R, *et al.* High temporal variability in commensal Escherichia coli strain communities of a herbivorous marsupial. *Environ Microbiol*. 2013;15:2162–72.

22 Massot M, Couffignal C, Clermont O, *et al.* Day-to-Day Dynamics of Commensal Escherichia coli in Zimbabwean Cows Evidence Temporal Fluctuations within a Host-Specific Population Structure. *Appl Environ Microbiol*. 2017;83. doi: 10.1128/AEM.00659-17
23 Reeves PR, Liu B, Zhou Z, *et al.* Rates of mutation and host transmission for an Escherichia coli clone over 3 years. *PloS One*. 2011;6:e26907.
24 Ghalayini M, Launay A, Bridier-Nahmias A, *et al.* Evolution of a Dominant Natural Isolate of Escherichia coli in the Human Gut over the Course of a Year Suggests a Neutral Evolution with Reduced Effective Population Size. *Appl Environ Microbiol*. 2018;84:e02377-17.
25 Wielgoss S, Barrick JE, Tenaillon O, *et al.* Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc Natl Acad Sci U S A*. 2013;110:222–7.
26 Tenaillon O, Barrick JE, Ribeck N, *et al.* Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*. 2016;536:165–70.
27 Bankevich A, Nurk S, Antipov D, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol J Comput Mol Cell Biol*. 2012;19:455–77.
28 Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl*. 2014;30:2068–9.
29 Touchon M, Hoede C, Tenaillon O, *et al.* Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet*. 2009;5:e1000344.
30 Burgaya J, Marin J, Royer G, *et al.* The bacterial genetic determinants of Escherichia coli capacity to cause bloodstream infections in humans. 2023;2022.12.31.522367. https://doi.org/10.1101/2022.12.31.522367
31 Petitjean M, Condamine B, Burdet C, *et al.* Phylum barrier and Escherichia coli intra-species phylogeny drive the acquisition of antibiotic-resistance genes. *Microb Genomics*. 2021;7. doi: 10.1099/mgen.0.000489
32 Bourrel AS, Poirel L, Royer G, *et al.* Colistin resistance in Parisian inpatient faecal Escherichia coli as the result of two distinct evolutionary pathways. *J Antimicrob Chemother*. 2019;74:1521–30.
33 Beghain J, Bridier-Nahmias A, Le Nagard H, *et al.* ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genomics*. 2018;4:e000192.
34 Joensen KG, Tetzschner AMM, Iguchi A, *et al.* Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol*. 2015;53:2410–26.
35 Ingle DJ, Valcanis M, Kuzevski A, *et al.* 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*. 2016;2:e000064.
36 Roer L, Tchesnokova V, Allesøe R, *et al.* Development of a Web Tool for Escherichia coli Subtyping Based on fimH Alleles. *J Clin Microbiol*. 2017;55:2538–43.
37 Seemann T. ABRicate. 2023. https://github.com/tseemann/abricate (accessed 26 April 2023)
38 Royer G, Decousser JW, Branger C, *et al.* PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. *Microb Genomics*. 2018;4:e000211.
39 Zankari E, Allesøe R, Joensen KG, *et al.* PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J Antimicrob Chemother*. 2017;72:2764–8.
40 Royer G, Darty MM, Clermont O, *et al.* Phylogroup stability contrasts with high within sequence type complex dynamics of Escherichia coli bloodstream infection isolates over a 12-year period. *Genome Med*. 2021;13:77.
41 Gautreau G, Bazin A, Gachet M, *et al.* PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput Biol*. 2020;16:e1007732.

42 Brynildsrud O, Bohlin J, Scheffer L, *et al.* Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol*. 2016;17:238.
43 Nguyen L-T, Schmidt HA, von Haeseler A, *et al.* IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32:268–74.
44 Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47:W256–9.
45 Tesson F, Hervé A, Mordret E, *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*. 2022;13:2561.
46 Camacho C, Coulouris G, Avagyan V, *et al.* BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
47 Wick RR, Judd LM, Gorrie CL, *et al.* Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput Biol*. 2017;13:e1005595.
48 Deatherage DE, Barrick JE. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol Clifton NJ*. 2014;1151:165–88.
49 Martinson JNV, Pinkham NV, Peters GW, *et al.* Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. *ISME J*. 2019;13:2306–18.
50 Fu YX. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol Biol Evol*. 2001;18:620–6.
51 Denamur E, Clermont O, Bonacorsi S, *et al.* The population genetics of pathogenic Escherichia coli. *Nat Rev Microbiol*. 2021;19:37–54.
52 Jacquier H, Assao B, Chau F, *et al.* Faecal carriage of extended-spectrum β-lactamase-producing Escherichia coli in a remote region of Niger. *J Infect*. 2023;S0163-4453(23)00336-5.
53 Le Gall T, Clermont O, Gouriou S, *et al.* Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group Escherichia coli Strains. *Mol Biol Evol*. 2007;24:2373–84.
54 Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol*. Published Online First: 17 July 2023. doi: 10.1038/s41579-023-00934-x
55 Armand-Lefèvre L, Rondinaud E, Desvillechabrol D, *et al.* Dynamics of extended-spectrum beta-lactamase-producing Enterobacterales colonization in long-term carriers following travel abroad. *Microb Genomics*. 2021;7. doi: 10.1099/mgen.0.000576
56 Ismail MD, Luo T, McNamara S, *et al.* Long-Term Carriage of Ciprofloxacin-Resistant Escherichia coli Isolates in High-Risk Nursing Home Residents. *Infect Control Hosp Epidemiol*. 2016;37:440–7.
57 Marin J, Clermont O, Royer G, *et al.* The Population Genomics of Increased Virulence and Antibiotic Resistance in Human Commensal Escherichia coli over 30 Years in France. *Appl Environ Microbiol*. 2022;88:e0066422.
58 Xu Y, Zhou N-Y. MhpA Is a Hydroxylase Catalyzing the Initial Reaction of 3-(3-Hydroxyphenyl)Propionate Catabolism in Escherichia coli K-12. *Appl Environ Microbiol*. 2020;86:e02385-19.
59 Kawano M. Divergently overlapping cis-encoded antisense RNA regulating toxin-antitoxin systems from E. coli: hok/sok, ldr/rdl, symE/symR. *RNA Biol*. 2012;9:1520–7.
60 Yang X, Tran F, Zhang P. Comparative Genomic Analyses of Escherichia coli from a Meat Processing Environment in Relation to Their Biofilm Formation and Persistence. *Microbiol Spectr*. 2023;11:e0018323.
61 Tanaka N, Shuman S. RtcB is the RNA ligase component of an Escherichia coli RNA repair operon. *J Biol Chem*. 2011;286:7727–31.
62 Cummins EA, Hall RJ, Connor C, *et al.* Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. *Microb Genomics*. 2022;8:mgen000903.
63 Hochhauser D, Millman A, Sorek R. The defense island repertoire of the Escherichia coli

pan-genome. *PLoS Genet*. 2023;19:e1010694.
64 Holgersen EM, Gandhi S, Zhou Y, *et al.* Transcriptome-Wide Off-Target Effects of Steric-Blocking Oligonucleotides. *Nucleic Acid Ther*. 2021;31:392–403.
65 Tal N, Millman A, Stokar-Avihail A, *et al.* Bacteria deplete deoxynucleotides to defend against bacteriophage infection. *Nat Microbiol*. 2022;7:1200–9.
66 Cui Y, Su X, Wang C, *et al.* Bacterial MazF/MazE toxin-antitoxin suppresses lytic propagation of arbitrium-containing phages. *Cell Rep*. 2022;41:111752.
67 Clermont O, Christenson JK, Denamur E, *et al.* The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep*. 2013;5:58–65.
68 Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*. 2000;28:292.
69 Lee H, Popodi E, Tang H, *et al.* Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A*. 2012;109:E2774-2783.
70 Touchon M, Perrin A, de Sousa JAM, *et al.* Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLoS Genet*. 2020;16:e1008866.
71 Salzberg SL. Open questions: How many genes do we have? *BMC Biol*. 2018;16:94.
72 Smati M, Clermont O, Le Gal F, *et al.* Real-time PCR for quantitative analysis of human commensal Escherichia coli populations reveals a high frequency of subdominant phylogroups. *Appl Environ Microbiol*. 2013;79:5005–12.
73 Nowrouzian FL, Wold AE, Adlerberth I. Escherichia coli strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *J Infect Dis*. 2005;191:1078–83.
74 Mäklin T, Thorpe HA, Pöntinen AK, *et al.* Strong pathogen competition in neonatal gut colonisation. *Nat Commun*. 2022;13:7417.
75 Morel-Journel T, Lehtinen S, Cotto O, *et al.* Persistence-colonization trade-off and niche differentiation enable the coexistence of E. coli phylogroups. 2023;2023.08.10.552787. https://doi.org/10.1101/2023.08.10.552787
76 Picard B, Garcia JS, Gouriou S, *et al.* The link between phylogeny and virulence in Escherichia coli extraintestinal infection. *Infect Immun*. 1999;67:546–53.
77 Massot M, Daubié A-S, Clermont O, *et al.* 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. *Microbiol Read Engl*. 2016;162:642–50.
78 Ludden C, Raven KE, Jamrozy D, *et al.* One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. *mBio*. 2019;10:e02693-18.
79 Berthe T, Ratajczak M, Clermont O, *et al.* Evidence for coexistence of distinct Escherichia coli populations in various aquatic environments and their survival in estuary water. *Appl Environ Microbiol*. 2013;79:4684–93.
80 Chang D-E, Smalley DJ, Tucker DL, *et al.* Carbon nutrition of Escherichia coli in the mouse intestine. *Proc Natl Acad Sci*. 2004;101:7427–32.
81 Burlingame R, Chapman PJ. Catabolism of phenylpropionic acid and its 3-hydroxy derivative by Escherichia coli. *J Bacteriol*. 1983;155:113–21.
82 Luria SE, Human ML. A nonhereditary, host-induced variation of bacterial viruses. *J Bacteriol*. 1952;64:557–69.
83 Barrangou R, Fremaux C, Deveau H, *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 2007;315:1709–12.
84 Touchon M, Charpentier S, Clermont O, *et al.* CRISPR distribution within the Escherichia coli species is not suggestive of immunity-associated diversifying selection. *J Bacteriol*. 2011;193:2460–7.
85 Doron S, Melamed S, Ofir G, *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*. 2018;359:eaar4120.
86 Stokar-Avihail A, Fedorenko T, Hör J, *et al.* Discovery of phage determinants that confer

sensitivity to bacterial immune systems. *Cell*. 2023;186:1863-1876.e16.
87 Hsu BB, Gibson TE, Yeliseyev V, *et al.* Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. *Cell Host Microbe*. 2019;25:803-814.e5.
88 Royer G, Clermont O, Marin J, *et al.* Epistatic interactions between the high pathogenicity island and other iron uptake systems shape Escherichia coli extra-intestinal virulence. *Nat Commun*. 2023;14:3667.
89 Baas Becking LGM. *Geobiologie of inleiding tot de milieukunde*. Den Haag: W.P. Van Stockum & Zoon 1934.
90 Bridier-Nahmias A, Launay A, Bleibtreu A, *et al.* Escherichia coli Genomic Diversity within Extraintestinal Acute Infections Argues for Adaptive Evolution at Play. *mSphere*. 2021;6. doi: 10.1128/mSphere.01176-20
91 Levert M, Zamfir O, Clermont O, *et al.* Molecular and evolutionary bases of within-patient genotypic and phenotypic diversity in Escherichia coli extraintestinal infections. *PLoS Pathog*. 2010;6:e1001125.
92 Le Gac M, Plucain J, Hindré T, *et al.* Ecological and evolutionary dynamics of coexisting lineages during a long-term experiment with Escherichia coli. *Proc Natl Acad Sci U S A*. 2012;109:9487–92.
93 McNally A, Kallonen T, Connor C, *et al.* Diversification of Colonization Factors in a Multidrug-Resistant Escherichia coli Lineage Evolving under Negative Frequency-Dependent Selection. *mBio*. 2019;10:e00644-19.

Supplementary Online Material

Tables

Table S1. Main characteristics of the 210 *E. coli* ED isolates
Table S2. ST distribution of the 210 *E. coli* isolates gathered from the stools of the individual ED
Table S3. SNPs matrix of 210 *E. coli* ED isolates
Table S4. Main characteristics of the 35 *E. coli* ED clones
Table S5. Residence times of the 35 *E. coli* ED clones
Table S6. Comparative of softcore genomes of the 35 *E. coli* ED clones and the 359 *E. coli* control commensal strains
Table S7. List of mutations identified in nine *E. coli* ED clones corresponding to 128 isolates

Figures

Fig. S1. Maximum likelihood phylogenetic tree reconstructed from the core genome single nucleotide polymorphisms (SNPs) (n=3 217 191) of the 210 *E. coli* ED isolates.
Fig. S2. Mean of the cumulative number of replicons per sample of the 210 *E. coli* ED isolates over time.
Fig. S3. Mean number of resistant and virulence strains, and total number of defense systems in the 35 ED clones (vertical lines), together with the null distribution when sampling 35 strains at random among the 359 commensals of the control collection (histograms).
Fig. S4. Percentage of strains encoding each defense system between the ED clones and the control commensal collection.
Fig. S5. Core genome SNP phylogenetic tree of the ST399 Cplx strains from EnteroBase (n=113), ED clones (n=6) and commensal control collection (n=4) rooted on four closely related strains from the control collection.

Figures 1 to 5

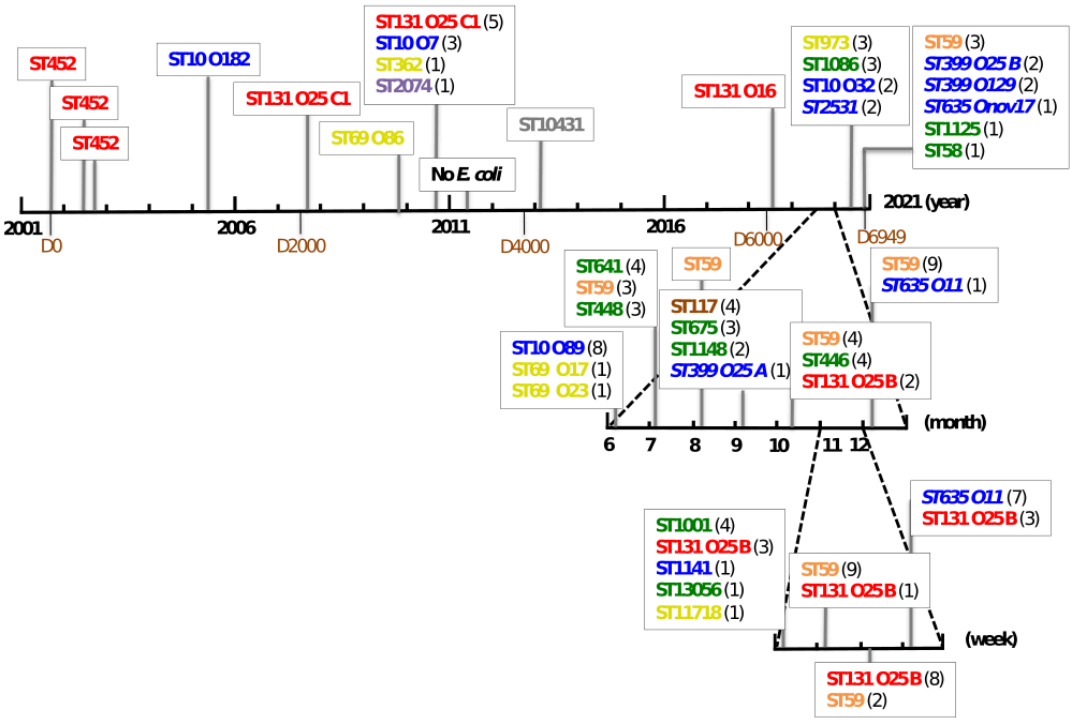

Fig. 1. Global representation of the study with the dates of sampling and the identified clones of the 22 stool samples performed during a 20 year-period. Ticks on the main (top) horizontal line correspond to one year intervals. Two periods with more frequent sampling are represented as sublines nested on the main line, with ticks defining monthly (middle line) and weekly (bottom line) intervals. Among the 22 samples, only one in 2011, did not yield any *E. coli* but a unique clone of *Hafnia alvei*. For the sake of clarity the *E. coli* clones are identified in a simplified way by their STs, and when necessary, their O-groups and their sub-types. The colors correspond to their phylogroup (A=blue, B1=green, B2=red, D=yellow, E=violet, F=orange, G=brown, H=grey). When the 10 isolates of a sample belong to the same clone, only one clone name is indicated. Otherwise, the number of isolates belonging to the clone is indicated in brackets. The first sample corresponds to day 0 in 2001 and the last one to day 6949 in 2020. The clones bearing the dGTPase and/or Detocs defense system(s) are indicated in italics.

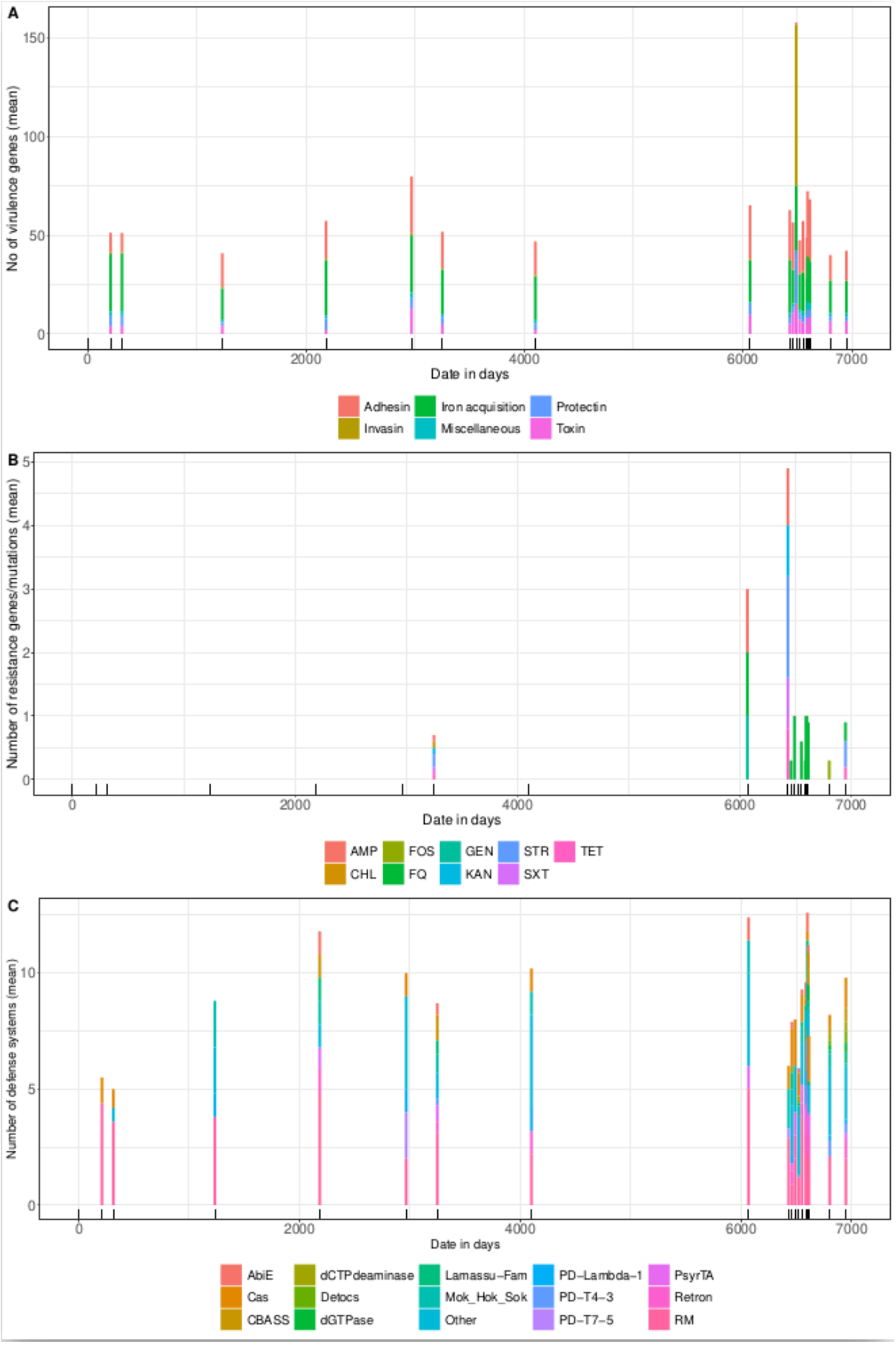
A
No of virulence genes (mean)
150
100
50
0
0
2000
4000
6000
7000
Date in days
Adhesin
Iron acquisition
Protectin
Invasin
Miscellaneous
Toxin
B
Number of resistance genes/mutations (mean)
5
4
3
2
1
0
0
2000
4000
6000
7000
Date in days
AMP
FOS
GEN
STR
TET
CHL
FQ
KAN
SXT
C
Number of defense systems (mean)
10
5
0
0
2000
4000
6000
7000
Date in days
AbiE
dCTPdeaminase
Lamassu−Fam
PD−Lambda−1
PsyrTA
Cas
Detocs
Mok_Hok_Sok
PD−T4−3
Retron
CBASS
dGTPase
Other
PD−T7−5
RM

Fig. 2. Mean of the cumulative number of (A) virulence associated genes, (B) antibiotic resistance and (C) antiviral systems per sample of the 210 *E. coli* isolates over time. At each date, the numbers of virulence associated genes, antibiotic resistance of each of the 10 isolates are summed and the mean per isolate presented. (A) The genes are colored by main functional categories. (B) The phenotypic resistance pattern of the strains to main classes of antibiotics was deduced from the genotype as in [40]. AMP: ampicillin, CHL: chloramphenicol, FOS: fosfomycin, FQ: fluoroquinolone, GEN: gentamycin, SXT: sulfamethoxazole trimethoprim, TET: tetracycline.

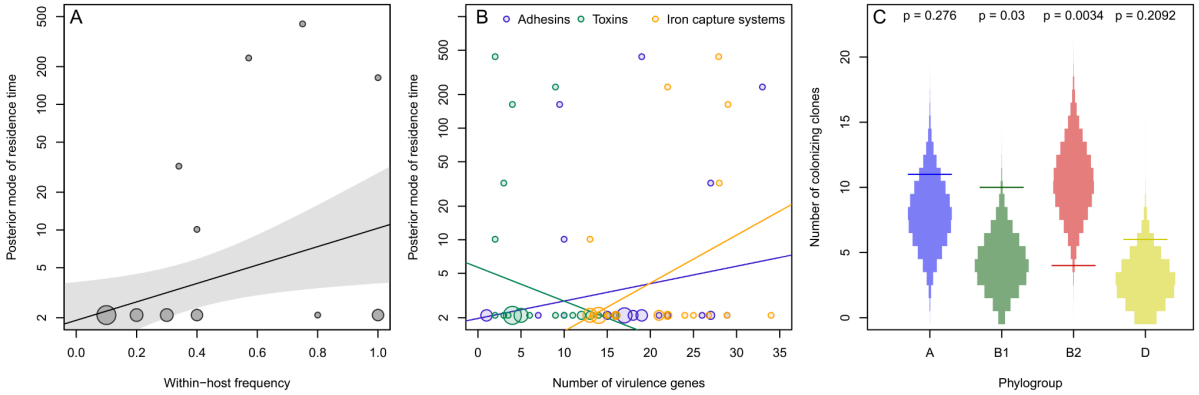

Fig. 3 Residence time and colonization rates of the *E. coli* ED clones. (A) Inferred posterior of residence time of clones as a function of its within-host frequency (circles, area proportional to the number of observations), with the predicted relationship by linear regression (solid line, with gray area representing the 95% confidence interval), (B) Inferred posterior of residence time of clones as a function of the number of virulence genes (circles, area proportional to the number of observations) associated with adhesins (purple), toxins (green) or iron capture systems (orange), with the relationships predicted by linear regression (solid line). (C) Observed number of colonizing clones for phylogroup A (blue), B1 (green), B2 (red) and D (yellow) (horizontal line). This is compared to the expectation if ED were in contact with the representative pool of *E. coli* in the population (the 359 healthy individuals), (colored areas), used to generate p-values corresponding to the probability of these values under this hypothesis.

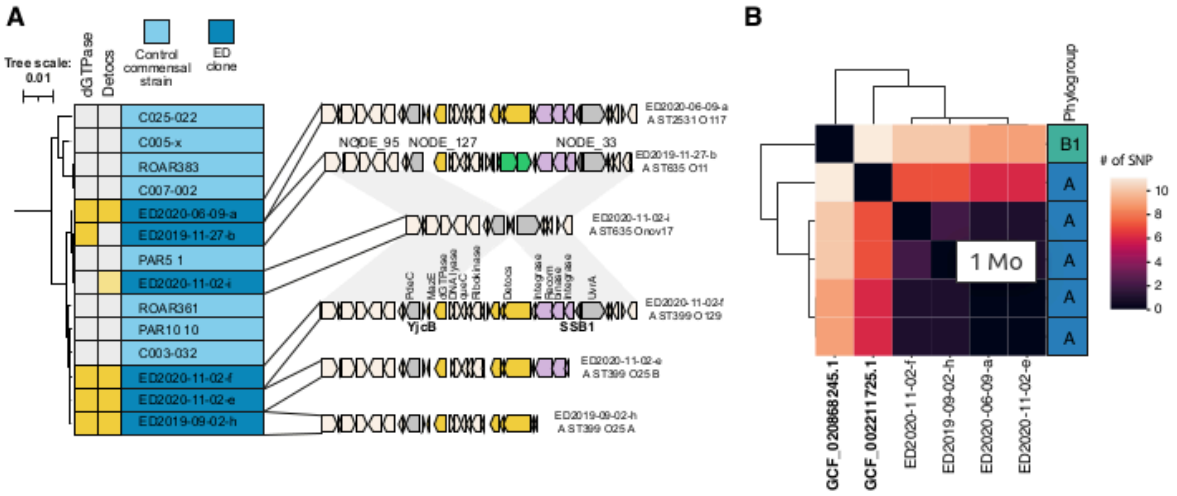

Fig. 4. *yjcB-ssb* antiviral defense island. (A) Schematic representation of the antiviral defense dGTPase and Detocs containing island found in complete and incomplete forms in four and one *E. coli* ED clones, respectively, with the synteny of the flanking regions in an isolate lacking this island (ED2020-11-02-i), but Detocs positive. Genes in yellow represent the different defense systems detected; in purple the mobility genes of the island; in green the insertion of another integrase inside the island; in gray the flanking genes *pdeC-yjcB* and *uvrA-ssb* genes. The island of the 2019-11-27-b isolate has been reconstructed using three different contigs (Nodes 95/127/33). The tree represents the phylogenetic history of the closely related phylogroup A0 strains belonging to the ED clone and control commensal collection based on the SNP core genome. Two main sub-lineages were observed, the upper branch corresponding to four island-negative strains whereas the lower branch corresponds to the Cplx399 with a mix of island-positive and negative strains. The presence or absence of dGTPase or Detocs is indicated in yellow. Of note, the ED2020-11-02-i isolate encodes for a different Detocs system (pale yellow), but in a different region of the genome and without the dGTPase. Conversely, the ED2019-11-27-b isolate is only dGTPase positive. (B). Clustermap based on the SNP matrix of the complete *yjcB-ssb* island (18 kbp) from the RefSeq database (in bold) and the ED clones. The A phylogroup strain from RefSeq belongs to the ST 635 whereas the B1 strain belongs to the ST58.

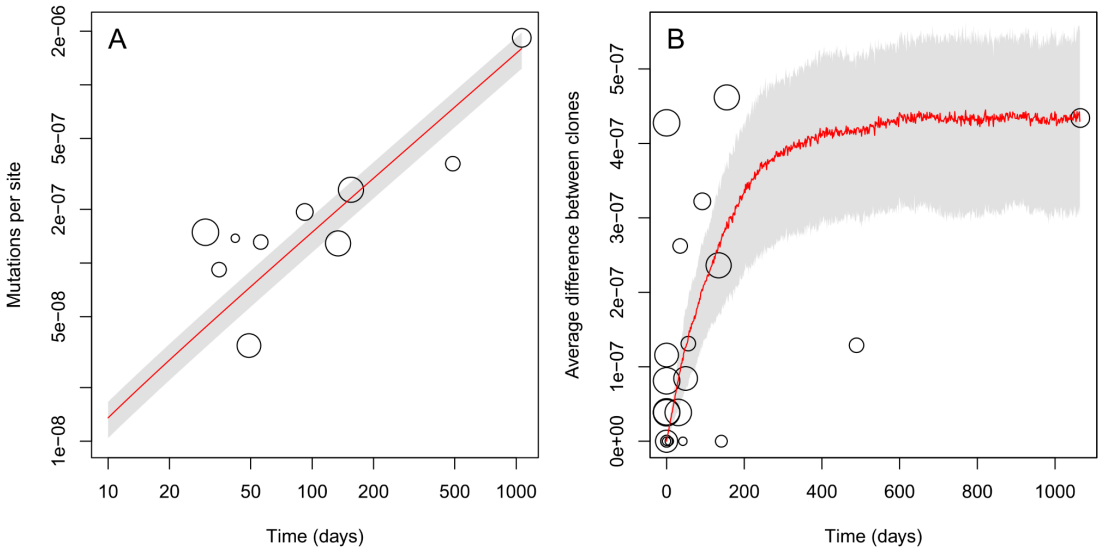

Fig. 5 Accumulation of mutations over time in the *E. coli* ED clones. (A) Number of mutations per site as a function of time elapsed since the first sampling (in log-scale), for each sampling of the four clones sampled multiple times after the first one (80 isolates, circles, size proportional to the number of isolates), with the estimated mutation rate (red line) with the predicted 95% confidence interval (gray). (B) Average difference between clones per site as a function of time since the first sampling, for each sampling of the nine clones considered (circles, size proportional to the number of isolates) (Table S7), and estimated according to 1,000 simulations of the Wright-Fisher model with Ne = 1000 (average over all simulations in red, interquartile range representing 50% of the simulations in gray).