Strain intrinsic properties and environmental constraints together shape *Escherichia coli* dynamics and diversity over a twenty-year human gut time series

Bénédicte Condamine[1], Thibaut Morel-Journel[1,2], Florian Tesson[1, 3], Guilhem Royer[1,4,5], Mélanie Magnan[6], Aude Bernheim[3], Erick Denamur[1,7*], François Blanquart[8*], 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] Unité de Bactériologie, Département de Prévention, Diagnostic et Traitement des Infections, AP-HP, Hôpital Henri Mondor, Créteil, France.

[5] EA 7380 Dynamyc, EnvA, UPEC, University of Paris-Est, Créteil, France

[6] Université Paris Cité, INSERM, CNRS, Institut Cochin, UMR 1016, 75014, Paris, 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

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

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

All authors declare there are no competing financial interests in relation to the work described.

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 commensal niche, the vertebrate gut. Using Illumina and/or Nanopore technologies, we sequenced whole genomes of 210 *E. coli* isolates from 22 stools sampled during a 20-year period from a healthy man (ED) living in Paris, France. All phylogroups, except C, were represented, with a predominance of B2 (34.3%), followed by A and F (19% each) phylogroups. Thirty-five clones were identified based on their haplogroup and pairwise genomic single nucleotide polymorphism distance and classified in three phenotypes according to their abundance and residence time: 25 sub-dominant/transient (52 isolates), five dominant/transient (48 isolates) and five dominant/resident (110 isolates). Four over five dominant/resident clones belonged to B2 and closely related F phylogroups, whereas sub-dominant/transient clones belonged mainly to B1, A and D phylogroups. The long residence times of B2 clones seemed to be counterbalanced by lower colonization abilities. Clones with larger within-host frequency persisted for longer in the host. By comparing ED strain genomes to a collection of commensal *E. coli* genomes from 359 French individuals, we identified ED-specific genomic properties including a set of genes involved in a metabolic pathway (*mhp* cluster) and a very rare antiviral defense island. The *E. coli* colonization within the gut microbiota was shaped by both the intrinsic properties of the strain lineages, in particular longer residence of phylogroup B2, and the environmental constraints such as diet or phages.

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). 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 for a large part in the gut, as a commensal (5). Strikingly, despite the urgent need to understand the normal ecology of *E. coli* to fight or manage 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 dominant commensal isolates are still rare compared to clinical ones.

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 over a few months (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 colonization ability: those persisting for short periods (days to weeks), the “transients” and those persisting for longer periods (but not on very long timescales), 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 an observed mutation rate per base per year of 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, we extended the study of ED *E. coli* carriage 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

Ethics

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). The authors were given consent by the subject.

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 (see supplementary online material (SOM) for more details on ED), 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 without attempting to maximize phenotypic diversity (see SOM) and stored at – 80°C in glycerol (220 isolates for 22 samples). The isolates of the three first samplings have been published previously (24).

Genome sequencing

All ED isolates were sequenced using Illumina technology (short reads) on MiniSeq sequencer. The 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 *E. coli* 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 *E. coli* genomes of high quality were downloaded from NCBI (July 2022).

Genome analyses

Genome typing was performed as previously described (31) (https://github.com/iame-researchCenter/Petanc), including species identification, phylogrouping (32), multi-locus sequence type (MLST) determination according to the Warwick University and Pasteur Institute schemes, *in silico* serotyping (33,34) and *fimH* allele determination (35). The presence of virulence associated genes (coding for adhesins, iron capture systems, toxins, protectins) and resistance genes was determined as previously

described with ABRicate 1.0.0, based on a custom database and Resfinder, respectively (36). The chromosomal or plasmidic location of a gene was determined using PlaScope 1.3.1 (37). We also searched for point mutations responsible for resistance using PointFinder 4.2.2 (38). 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 (39).

Pan and core genomes were computed using PPanGGOLiN 1.2.105 with default parameters (40). Pairwise single nucleotide polymorphism (SNP) distance matrices from multiple sequence alignments were performed using pairsnp (https://github.com/gtonkinhill/pairsnp). Clones were defined according to the following strategy (41). First, isolates were grouped by their haplogroup, defined as a combination of their sequence types (ST) according to the Warwick University and Institute Pasteur schemes, their serotype (O:H), and their *fimH* allele. Then, the distribution of core genome pairwise SNP distances was determined to identify a cut-off where few/no values were observed, that can be used as a threshold to define clones. The SNP distances between isolates were compared with their haplogroup, allowing a data-based definition of the clone.

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 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). Additional methods to analyze the genetics of the defense island and *mhp* cluster are detailed in the SOM.

Illumina and nanopore contigs were assembled using Unicycler v0.4.9b (46) and annotated with Prokka 1.14.5 (28). We then filtered the mutations identified by breseq 0.33.2 (47) 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. The exponential distribution is a simple, one-parameter distribution that emerges when the clearance rate is constant in time; we chose it for its simplicity in the absence of detailed data that could inform on potentially more complex distributions. To use prior information on residence times detailed by published studies, we adopted a Bayesian inference approach. 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 (48).

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

We studied the within-clone diversity, based on the mutations accumulated by each clone during its residence in the ED (see SOM for details on the method). On the one hand, we estimated an observed mutation rate based on differences between isolates of a given clone. On the other hand, we compared the observed differences to simulations of the Wright-Fisher model to assess the expected effective population size in ED under the hypothesis of neutrality.

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). Except for one sample from 2011 which yielded 10 isolates of a unique clone of *Hafnia alvei*, all other samples were positive for *E. coli stricto sensu* (49), leading to a collection of 210 *E. coli* isolates (Table S1).

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

ED isolate core and pan genomes

The soft core genome (hereafter, "core") of the 210 isolates (*i.e.*, genes present in 95% to 100% of the isolates) was composed of 3,278 genes. On the other hand, the pangenome (the totality of genes) encompassed 13,627 genes. The core and pangenome of the clones were similar to that of the isolates, although slightly smaller (3,270 and 12,986, respectively).

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 (49), 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)	/

For the delineation of clones, thirty four distinct haplogroups based on a combination of known typing markers (see methods for precise definition) were first identified (41). Then, the pairwise core genome SNP distance distribution was analyzed and confronted to the haplogroup of the isolates. It clearly appeared a cutoff separating inter-haplogroup isolates differing by more than 1840 SNPs from intra-haplogroup isolates differing by 21 SNPs or less (Table S3, Fig. S2). The only exception was in the phylogroup A_ ST399/661_O25:H12_fimHND haplogroup, where isolates differed by 125 to 130 SNPs. We

checked the integrity of the mismatch repair genes in these isolates to exclude mutators (50). Based on our data, together with a recent modeling study defining genomic epidemiology thresholds from bacterial outbreaks between one to 19 SNPs (51), we considered that this haplogroup encompassed two different clones, named A and B. Nevertheless, we cannot exclude that all the isolates of this haplogroup belong to the same clone that evolved in the ED gut. In sum, our data-driven approach identified a maximum of 35 clones named thereafter "ED clones" among the 210 *E. coli* isolates (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).

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. S3).

We then focused on specific genetic elements known to be important for *E. coli* commensal lifestyle in the gut: the virulence associated genes, as it has been proposed that virulence is a by-product of commensalism (52), the antibiotic resistance genes, as antibiotics are major selective agents and the defense islands against phage predation (53). These factors may be primarily colonization factors (4). We also compared the ED clones to an epidemiologically relevant control collection of 359 *E. coli* isolates from the stools of 359 healthy individuals living in France during the same period (54), 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 associated genes than the commensals (mean 47 in ED against 56 in commensals, $p = 0.00014$ according to a permutation test) (Fig. S4).

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. S4).

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. S4). 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. S5). We will come back to this finding below in a more systematic analysis.

Epidemiological dynamics of ED clones

We then investigated 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.

Residence 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 (48). We inferred the posterior distribution of the residence time for each observed clone (Table S4). Most of the clones (30)

out of 35) were seen only once and consequently had very wide inferred posterior residence times distributions (with a mode at 2.1 days), reflecting the poor information on the true residence times of these clones in our data. The remaining five clones classified as resident have a posterior mode of residence time 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. But most (15 over 21) timepoints have a dominant clone.

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 (110 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 (24.7)	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 (22.8)	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 (52.4)	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.

Long term intestinal residence of colonizing *E. coli* has been associated with the presence of specific clones belonging mainly to B2/D/F phylogroups, iron capture systems (17,55), and fluoroquinolone resistance (56). We did a series of tests for the association between the clone category, or inferred posterior mode of residence time, and various genetic factors. Although we found that clones with longer residence time exhibited more iron capture systems (Table 2), this association disappeared when controlling for the phylogroup.

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

The epidemiological dynamics of *E. coli* rely not only on the intestinal residence of clones, but also on their ability to colonize new hosts, 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 (54). Colonizations of the gut of ED by B1 and B2 phylogroups were respectively more frequent (4.72 more events than expected, $p = 0.0301$) and rarer (7.5 less events than expected, $p = 0.0034$), suggesting intrinsic differences in colonization rates (Fig. 3B). It therefore appears that the high residence time of B2 clones is counterbalanced by a lower colonization ability.

Genome-wide genomic features associated with clones successfully colonizing ED

To identify distinct genetic features of *E. coli* strains able to colonize the focal individual, we performed a systematic genome-wide association study (GWAS) using Scoary (42) to

investigate if some genes were differentially retrieved between the 35 ED pangenomes and the pangenomes from the 359 *E. coli* commensal strains of the control collection (54). The core genome of this control collection was composed of 3,262 genes whereas the pangenome encompassed 27,752 genes.

Indeed, we found 36 genes significantly associated with the 35 ED clones, of which 32 are enriched (rather than depleted) in ED clones (Table S5). Two clusters of biologically relevant genes were enriched in ED clones: a cluster of eight genes (*mhp*) allowing the use of the 3-(3-hydroxyphenyl)propionate (3HPP), one of the phenylpropanoids from lignin (57) and a cluster of 10 genes found in a phage defense island. In depth analysis of these two clusters is presented below.

mhp *cluster*

This cluster is composed of the transcriptional activator gene (*mhpR*), enzyme coding genes *mhpA* to *F* and transport protein gene *mhpT* flanked by two bacterial interspersed mosaic element (BIME) (58). It was found in all 35 ED clones but only in 71.6% (257/359) of the control strains. An association with specific STs has been recently reported for this gene cluster (59), which is present in all phylogroups but not B2 where it is absent except in ST131. 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.

To elucidate the phylogenetic scenario leading to such pattern, we searched for this cluster in 1,777 *E. coli* genomes of high quality from the RefSeq database, reconstructed the *mhp*

phylogenetic history and compared it to the strain phylogeny. We first confirmed the presence of this cluster in all phylogroups except B2, where it is present in the ST131 and some single locus variants (ST131-like) (Fig. 4A and Table S6). We also found a *mhp* positive strain related to ST452 (ST3693, double locus variant), and six other *mhp* positive B2 strains from various and rare STs (ST646, ST7898, ST91, ST136, ST681, ST1155) scattered across the B2 phylogenetic tree (Fig. 4B). The entire *mhp* cluster phylogeny (Fig. 4A) is congruent with the strain phylogeny except for B2 strains (Fig. 4B). The *mhp* found in the latter can be divided into three main groups: the *mhp*-B2-a which gathers three strain pairs distributed throughout the tree; the *mhp*-B2-b found in highly divergent STs (ST646 and ST7898); the *mhp*-B2-c carried by ST131 strains (Fig. 4B). The physical map of the *mhp* cluster with flanking regions of representative *mhp* positive strains from phylogroup B2 (Fig. S6) showed an almost perfect synteny, i.e. conservation of genes order on the chromosome. Insertion sequences (IS*1,* IS*3* and IS*110*), are found in few genomes. But given their location, their diversity and the highly conserved position of the *mhp* gene cluster, we can probably rule out any role in the acquisition of this determinant.

These data argue for an ancient arrival of the *mhp* cluster in the species followed by its loss in the ancestral B2 strain and several independent re-acquisitions probably by homologous recombination of flanking DNA regions as previously described for the High-Pathogenicity Island (HPI) (60). Another scenario, albeit less parsimonious, would be the presence of the cluster in the ancestral B2 strain, followed by multiple independent deletion events.

Antiviral defense island

A cluster composed of 15 genes corresponding to an antiviral defense island (45), was present in five (5/35, 14.3%) ED clones (in complete and incomplete form in four and one clones, respectively) but never in the 359 control strains (Fig. 5A, Table S5 - 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 (61). 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) (62). (ii) a *dGTPase* gene that encodes deoxyguanosine triphosphatase (dGTPase) enzyme, depleting the nucleotide pool and halting the phage replication (63). This gene is adjacent to (iii) *mazE*, coding for the antitoxin of the MazF/MazE module that suppresses phage lytic propagation via abortive infection (64). 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 -) (65). They are intertwined with the ED2020-11-02-i isolate and with island-negative control strains (Fig. 5A). When searching for this island in the *E. coli* genomes from the RefSeq database, we found only two positive strains for the complete island belonging to phylogroup/ST A ST635 and B1 ST58

(Fig. 5B). SNP comparison of 12 kbp 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. 5B), 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 9-11 SNPs).

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.

Within-host evolution of resident ED clones: no evidence for diversifying selection and estimation of the 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 (47) 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 (66) including more than 1,000 codons. The average ratio was 0.836 (from 0.730 to 0.889), consistent with weak purifying selection.

The observed 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 (67), 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. 6A). This observed rate does not include the highly deleterious mutations which cannot be observed in our approach. Based on previous estimations of the mutation rate at 2.2×10^{-10} mutation per generation per nucleotide (67), 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 (N_e) 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 genotypes over time, for the estimated mutation rate and various effective population sizes. The divergence observed in ED was consistent with a value of N_e of the order of 1000 bacteria, the model predicting an equilibrium reached in about a year (Fig 6B). For lower effective sizes (e.g. $N_e = 100$), the divergence stabilized at a much lower equilibrium than observed, while for larger ones (e.g. $N_e = 10000$), it continued increasing after a year towards much higher values.

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 additionally 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) (68) (data not shown). The pangenome was composed of more than 13,000 genes. Considering that the human genome encompasses 20,000 protein-coding genes (69), 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/residence 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 (70).

The pattern of coexistence of phylogroups within the single ED subject 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 (71). 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 (49).

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 (72). 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 (48). 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 (73). A re-analysis of the data from Martinson and Walk (48) and a study on a cohort of 46 healthy French adults (70) showed that the long residence of some

phylogroups was counterbalanced by a poorer capacity to colonize hosts. This residence-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 (70).

As there is a link between B2 strains and the abundance of virulence genes (74), 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 residence. However, in the multivariate analysis, multiple virulence genes (especially the iron uptake systems) and fluoroquinolones were better predictors of residence times than the ST131 background. Our data, possibly because of the small sample size, were also non conclusive. Larger datasets on the duration of residence 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 and D 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%) (75), B1 strains being commonly found in cattle (76) and also in water and sediments (77).

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. These

genes are involved in two fundamental processes of gut colonization: metabolism (78) and anti phage defense (53). They are localized on mobile genetic elements and selected together with the clones bearing them.

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 (79). It is tempting to speculate that the enrichment in *mhp* genes in ED is linked to his high and daily consumption of cereals (see SOM). Further microbiota studies should be performed in relation to diet to test the hypothesis. The *mhp* cluster, invading the *E. coli* species probably by horizontal gene transfer (58), is adjacent to the lactose operon (Fig. S6), both sets of genes allowing catabolism of breakfast compounds, i.e. milk and cereals.

Bacteria protect themselves from foreign DNA as phage DNA by more than a hundred systems, unveiling many different molecular mechanisms (53). 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, suggesting that these clones are selected by 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) may be involved is available in the literature (80). Moreover, the rarity of this island suggests strong selection 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" (81). Individual hosts constantly ingest large numbers of diverse 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 major sign of selection 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 (82). At the opposite, no molecular footprint of selection has been identified in commensal *E. coli* isolates from healthy individuals (83), including the previously published ST452 clone isolated from ED (24). In this study, we did not evidence a strong pattern of diversifying selection as in extra-intestinal sites, but weak purifying counter-selection of deleterious mutations. Nevertheless, we note that one clone splitted into two haplotypes separated by six mutations (see SOM), possibly by diversifying selection, but the wider significance of this single observed event was unclear. We acknowledge though that clear signatures of within-host selection could possibly be detected in larger datasets identifying more resident strains.

Besides, we confirmed that levels of diversity are compatible with an effective population size much lower than census size, possibly due to population structure within the gut.

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. Although we evidenced selection at the between-host level, no major sign of within-host molecular adaptation was identified. Such long term time series studies of *E. coli* microbiomes in a large range of subjects worldwide should be supported to generalize our findings.

Acknowledgments

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

Funding

B.C., E.D. and O.C. were partially supported by the "Fondation pour la Recherche Médicale" (Equipe FRM 2016, grant number DEQ20161136698). F.T. and A.B. were supported by the ATIP-Avenir program from INSERM (R21042KS/RSE22002KSA), ERC Starting Grant (PECAN 101040529). F.B. was funded by the ERC Starting Grant 949208 EvoComBac.

Competing Interests

All authors declare there are no competing financial interests in relation to the work described.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available in ENA database on the bioproject PRJEB45090 (https://www.ebi.ac.uk/ena/browser/view/PRJEB45090).

References

1. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, 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 Nov;18(11):1229–40.
2. Feldman SF, Temkin E, Wullfhart L, Nutman A, Schechner V, Shitrit P, 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 Jun;28(6):879.e1-879.e7.
3. Woerther PL, Burdet C, Chachaty E, Andremont A. Trends in human fecal carriage of extended-spectrum β-lactamases in the community: toward the globalization of CTX-M. Clin Microbiol Rev. 2013 Oct;26(4):744–58.
4. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal Escherichia coli. Nat Rev Microbiol. 2010 Mar;8(3):207–17.
5. Tedijanto C, Olesen SW, Grad YH, Lipsitch M. Estimating the proportion of bystander selection for antibiotic resistance among potentially pathogenic bacterial flora. Proc Natl Acad Sci U S A. 2018 Dec 18;115(51):E11988–95.
6. Martinson JNV, Walk ST. Escherichia coli Residency in the Gut of Healthy Human Adults. EcoSal Plus. 2020 Sep;9(1).
7. Totsuka K. Studien über Bacterium coli. Z Für Hyg Infekt. 1903 Dec 1;45(1):115–24.
8. Wallick H, Stuart CA. Antigenic Relationships of Escherichia coli Isolated from One Individual. J Bacteriol. 1943 Feb;45(2):121–6.
9. Perch B. Weitere Untersuchungen über Die Coliflora Des Gesunden Menschen. Acta Pathol Microbiol Scand. 1944;21(2):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 Jan;63(1):47–57.
11. Robinet HG. Relationship of host antibody to fluctuations of Escherichia coli serotypes in the human intestine. J Bacteriol. 1962 Nov;84(5):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 Jul;98(3):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 Dec;107(3):537–42.
14. McBurney W, Mccartney A, Apun K, Mcconnell M, Tannock G. Perturbation of the enterobacterial microflora detected by molecular analysis. Microb Ecol Health Dis. 1999 Sep 1;11.
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 Dec;46(12):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 Oct;137(10):1457–64.
17. Johnson JR, Clabots C, Porter SB, Bender T, Johnston BD, Thuras P. Intestinal Persistence of Colonizing Escherichia coli Strains, Especially ST131-H30, in Relation to Bacterial and Host Factors. J Infect Dis. 2022 Jun 15;225(12):2197–207.
18. Hu D, Fuller NR, Caterson ID, Holmes AJ, Reeves PR. Single-gene long-read sequencing illuminates Escherichia coli strain dynamics in the human intestinal microbiome. Cell Rep. 2022 Jan 11;38(2):110239.
19. Han N, Peng X, Zhang T, Qiang Y, Li X, Zhang W. Rapid turnover and short-term blooms of Escherichia coli in the human gut. J Bacteriol. 2024 Jan 25;206(1):e0023923.
20. Sears HJ, Brownlee I, Uchiyama JK. Persistence of individual strains of Escherichia coli in the intestinal tract of man. J Bacteriol. 1950 Feb;59(2):293–301.
21. Blyton MDJ, Banks SC, Peakall R, Gordon DM. High temporal variability in commensal Escherichia coli strain communities of a herbivorous marsupial. Environ Microbiol. 2013 Aug;15(8):2162–72.
22. Massot M, Couffignal C, Clermont O, D'Humières C, Chatel J, Plault N, 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 Jul 1;83(13).

23. Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, et al. Rates of mutation and host transmission for an Escherichia coli clone over 3 years. PloS One. 2011;6(10):e26907.
24. Ghalayini M, Launay A, Bridier-Nahmias A, Clermont O, Denamur E, Lescat M, 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 Mar 15;84(6):e02377-17.
25. Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, et al. Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci U S A. 2013 Jan 2;110(1):222–7.
26. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature. 2016 Aug 11;536(7615):165–70.
27. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol. 2012 May;19(5):455–77.
28. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma Oxf Engl. 2014 Jul 15;30(14):2068–9.
29. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009 Jan;5(1):e1000344.
30. Burgaya J, Marin J, Royer G, Condamine B, Gachet B, Clermont O, et al. The bacterial genetic determinants of Escherichia coli capacity to cause bloodstream infections in humans. PLoS Genet. 2023 Aug;19(8):e1010842.
31. Bourrel AS, Poirel L, Royer G, Darty M, Vuillemin X, Kieffer N, et al. Colistin resistance in Parisian inpatient faecal Escherichia coli as the result of two distinct evolutionary

pathways. J Antimicrob Chemother. 2019 Jun 1;74(6):1521–30.

32. Beghain J, Bridier-Nahmias A, Le Nagard H, Denamur E, Clermont O. ClermonTyping: an easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. Microb Genomics. 2018 Jul;4(7):e000192.
33. Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol. 2015 Aug;53(8):2410–26.
34. Ingle DJ, Valcanis M, Kuzevski A, Tauschek M, Inouye M, Stinear T, 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 Jul;2(7):e000064.
35. Roer L, Tchesnokova V, Allesøe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, et al. Development of a Web Tool for Escherichia coli Subtyping Based on fimH Alleles. J Clin Microbiol. 2017 Aug;55(8):2538–43.
36. Seemann T. ABRicate [Internet]. 2023 [cited 2023 Apr 26]. Available from: https://github.com/tseemann/abricate
37. Royer G, Decousser JW, Branger C, Dubois M, Médigue C, Denamur E, et al. PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the species level. Microb Genomics. 2018 Sep;4(9):e000211.
38. Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother. 2017 Oct 1;72(10):2764–8.
39. Royer G, Darty MM, Clermont O, Condamine B, Laouenan C, Decousser JW, 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 May 5;13(1):77.
40. Gautreau G, Bazin A, Gachet M, Planel R, Burlot L, Dubois M, et al. PPanGGOLiN:

Depicting microbial diversity via a partitioned pangenome graph. PLoS Comput Biol. 2020 Mar;16(3):e1007732.

41. Jacquier H, Assao B, Chau F, Guindo O, Condamine B, Magnan M, et al. Faecal carriage of extended-spectrum β-lactamase-producing Escherichia coli in a remote region of Niger. J Infect. 2023 Jun 25;S0163-4453(23)00336-5.
42. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016 Nov 25;17(1):238.
43. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.
44. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019 Jul 2;47(W1):W256–9.
45. Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun. 2022 May 10;13(1):2561.
46. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLOS Comput Biol. 2017 Jun 8;13(6):e1005595.
47. 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.
48. Martinson JNV, Pinkham NV, Peters GW, Cho H, Heng J, Rauch M, et al. Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. ISME J. 2019 Sep;13(9):2306–18.
49. Denamur E, Clermont O, Bonacorsi S, Gordon D. The population genetics of pathogenic Escherichia coli. Nat Rev Microbiol. 2021 Jan;19(1):37–54.
50. Denamur E, Matic I. Evolution of mutation rates in bacteria. Mol Microbiol. 2006

May;60(4):820–7.

51. Duval A, Opatowski L, Brisse S. Defining genomic epidemiology thresholds for common-source bacterial outbreaks: a modelling study. Lancet Microbe. 2023 May;4(5):e349–57.
52. Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, et al. Extraintestinal Virulence Is a Coincidental By-Product of Commensalism in B2 Phylogenetic Group Escherichia coli Strains. Mol Biol Evol. 2007 Nov 1;24(11):2373–84.
53. Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. Nat Rev Microbiol. 2023 Jul 17;
54. Marin J, Clermont O, Royer G, Mercier-Darty M, Decousser JW, Tenaillon O, 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 Aug 9;88(15):e0066422.
55. Armand-Lefèvre L, Rondinaud E, Desvillechabrol D, Mullaert J, Clermont O, Petitjean M, et al. Dynamics of extended-spectrum beta-lactamase-producing Enterobacterales colonization in long-term carriers following travel abroad. Microb Genomics. 2021 Jul;7(7).
56. Ismail MD, Luo T, McNamara S, Lansing B, Koo E, Mody L, et al. Long-Term Carriage of Ciprofloxacin-Resistant Escherichia coli Isolates in High-Risk Nursing Home Residents. Infect Control Hosp Epidemiol. 2016 Apr;37(4):440–7.
57. Xu Y, Zhou NY. MhpA Is a Hydroxylase Catalyzing the Initial Reaction of 3-(3-Hydroxyphenyl)Propionate Catabolism in Escherichia coli K-12. Appl Environ Microbiol. 2020 Feb 3;86(4):e02385-19.
58. Díaz E, Ferrández A, Prieto MA, García JL. Biodegradation of aromatic compounds by Escherichia coli. Microbiol Mol Biol Rev MMBR. 2001 Dec;65(4):523–69, table of contents.
59. Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. Microb

Genomics. 2022 Nov;8(11):mgen000903.

60. Schubert S, Darlu P, Clermont O, Wieser A, Magistro G, Hoffmann C, et al. Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species. PLoS Pathog. 2009 Jan;5(1):e1000257.
61. Hochhauser D, Millman A, Sorek R. The defense island repertoire of the Escherichia coli pan-genome. PLoS Genet. 2023 Apr;19(4):e1010694.
62. Holgersen EM, Gandhi S, Zhou Y, Kim J, Vaz B, Bogojeski J, et al. Transcriptome-Wide Off-Target Effects of Steric-Blocking Oligonucleotides. Nucleic Acid Ther. 2021 Dec;31(6):392–403.
63. Tal N, Millman A, Stokar-Avihail A, Fedorenko T, Leavitt A, Melamed S, et al. Bacteria deplete deoxynucleotides to defend against bacteriophage infection. Nat Microbiol. 2022 Aug;7(8):1200–9.
64. Cui Y, Su X, Wang C, Xu H, Hu D, Wang J, et al. Bacterial MazF/MazE toxin-antitoxin suppresses lytic propagation of arbitrium-containing phages. Cell Rep. 2022 Dec 6;41(10):111752.
65. Clermont O, Christenson JK, Denamur E, Gordon DM. The Clermont Escherichia coli phylo-typing method revisited: improvement of specificity and detection of new phylo-groups. Environ Microbiol Rep. 2013 Feb;5(1):58–65.
66. Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic Acids Res. 2000 Jan 1;28(1):292.
67. Lee H, Popodi E, Tang H, Foster PL. 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 Oct 9;109(41):E2774-2783.
68. Touchon M, Perrin A, de Sousa JAM, Vangchhia B, Burn S, O'Brien CL, et al. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. PLoS Genet. 2020 Jun;16(6):e1008866.
69. Salzberg SL. Open questions: How many genes do we have? BMC Biol. 2018 Aug 20;16(1):94.

70. Morel-Journel T, Lehtinen S, Cotto O, Amia R, Clermont O, Dion S, et al. Residence-colonization trade-off and niche differentiation enable the coexistence of Escherichia coli phylogroups in healthy humans [Internet]. bioRxiv; 2024 [cited 2024 Jun 11]. p. 2023.08.10.552787. Available from: https://www.biorxiv.org/content/10.1101/2023.08.10.552787v2
71. Smati M, Clermont O, Le Gal F, Schichmanoff O, Jauréguy F, Eddi A, 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 Aug;79(16):5005–12.
72. 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 Apr 1;191(7):1078–83.
73. Mäklin T, Thorpe HA, Pöntinen AK, Gladstone RA, Shao Y, Pesonen M, et al. Strong pathogen competition in neonatal gut colonisation. Nat Commun. 2022 Dec 1;13(1):7417.
74. Picard B, Garcia JS, Gouriou S, Duriez P, Brahimi N, Bingen E, et al. The link between phylogeny and virulence in Escherichia coli extraintestinal infection. Infect Immun. 1999 Feb;67(2):546–53.
75. Massot M, Daubié AS, Clermont O, Jauréguy F, Couffignal C, Dahbi G, 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 Apr;162(4):642–50.
76. Ludden C, Raven KE, Jamrozy D, Gouliouris T, Blane B, Coll F, et al. One Health Genomic Surveillance of Escherichia coli Demonstrates Distinct Lineages and Mobile Genetic Elements in Isolates from Humans versus Livestock. mBio. 2019 Jan 22;10(1):e02693-18.
77. Berthe T, Ratajczak M, Clermont O, Denamur E, Petit F. Evidence for coexistence of distinct Escherichia coli populations in various aquatic environments and their survival in

estuary water. Appl Environ Microbiol. 2013 Aug;79(15):4684–93.

78. Chang DE, Smalley DJ, Tucker DL, Leatham MP, Norris WE, Stevenson SJ, et al. Carbon nutrition of Escherichia coli in the mouse intestine. Proc Natl Acad Sci. 2004 May 11;101(19):7427–32.
79. Burlingame R, Chapman PJ. Catabolism of phenylpropionic acid and its 3-hydroxy derivative by Escherichia coli. J Bacteriol. 1983 Jul;155(1):113–21.
80. Stokar-Avihail A, Fedorenko T, Hör J, Garb J, Leavitt A, Millman A, et al. Discovery of phage determinants that confer sensitivity to bacterial immune systems. Cell. 2023 Apr 27;186(9):1863-1876.e16.
81. Baas Becking LGM. Geobiologie of inleiding tot de milieukunde. Den Haag: W.P. Van Stockum & Zoon; 1934. 263 p.
82. Bridier-Nahmias A, Launay A, Bleibtreu A, Magnan M, Walewski V, Chatel J, et al. Escherichia coli Genomic Diversity within Extraintestinal Acute Infections Argues for Adaptive Evolution at Play. mSphere. 2021 Jan 6;6(1).
83. Levert M, Zamfir O, Clermont O, Bouvet O, Lespinats S, Hipeaux MC, et al. Molecular and evolutionary bases of within-patient genotypic and phenotypic diversity in Escherichia coli extraintestinal infections. PLoS Pathog. 2010 Sep 30;6(9):e1001125.

Figure legends

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). The number of isolates belonging to the clone is indicated in brackets. The first sample corresponds to day 0 (D0) in 2001 and the last one to day 6949 (D6949) in 2020. The clones bearing the dGTPase and/or Detocs defense system(s) are indicated in italics.

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 and antiviral systems of each of the 10 isolates are summed and the mean per isolate presented. (A) The genes are colored by main functional categories. Day 0 (D0) to day 6949 (D6949) are indicated in brown as in Fig. 1. (B) The phenotypic resistance pattern of the strains to main classes of antibiotics was deduced from the genotype as in (39). AMP: ampicillin, CHL: chloramphenicol, FOS: fosfomycin, FQ: fluoroquinolone, GEN: gentamycin, SXT: sulfamethoxazole trimethoprim, TET: tetracycline. (C) Antiviral systems colored by functional groups.

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) 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 was 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. Phylogeny of the *mhp* gene cluster and its distribution among the phylogroup B2 phylogenetic history. (A) Maximum likelihood phylogenetic tree of the *mhp* gene cluster (7,884 pb) detected in 1576 over 1777 tested RefSeq genomes. The ED1a (NC_011745.1) *mph* gene cluster has also been included as no ST452 is available in the RefSeq dataset. The *mhp* carried by B2 strains is divided into three groups named *mph*-B2-a, *mph*-B2-b and *mph*-B2-c. The phylogroups are highlighted in color and the STs for the genomes belonging to phylogroup B2 are detailed. The tree is rooted on *Escherichia* clade I sequences. (B) Maximum likelihood phylogenetic tree computed from core gene alignment of the genomes belonging to phylogroup B2 in RefSeq genomes (n=282/1777). The genome of ED1a (NC_011745.1) has also been included. The presence of the *mph* gene cluster is highlighted by triangles colored according to the *mhp* groups previously defined: *mph*-B2-a (green), *mph*-B2-b (blue) and *mph*-B2-c (red). The tree is rooted on a phylogroup G ST117 genome (GCF_001900735.1). The genomes labeled in red correspond to the genomes used to analyse the genetic environment of *mhp* presented Fig. S6. Tree scales represent the number of nucleotide substitutions per site.

Fig. 5. The *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. 6 Accumulation of mutations over time in the *E. coli* ED clones. (A) Number of observed mutations per site (in log-scale) 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 observed mutation rate (red line) with the predicted 95% confidence interval (gray). (B) Average difference between isolates of the same clone per site (in log-scale), as a function of

time since the first sampling (in log-scale), for each sampling of the nine clones considered (circles, size proportional to the number of isolates) (Table S7), and averaged over 1,000 simulations of the Wright-Fisher model with N_e = 100 (black line), with N_e = 1000 (red line with interquartile range representing 50% of the simulations in gray) and with N_e = 10000 (blue line).

Supplementary online material (SOM)

Supplementary Tables

Table S1. Main characteristics of the 210 *E. coli* ED isolates

Table S2. ST distribution of the 210 *E. coli* ED isolates

Table S3. SNPs matrix of 210 *E. coli* ED isolates

Table S4. Main characteristics of the 35 *E. coli* ED clones

Table S5. List of genes significantly associated to the 35 ED clones

Table S6. Distribution of *mhp* gene cluster among 1,777 complete *E. coli* genomes from RefSeq

Table S7. List of mutations identified in nine *E. coli* ED clones corresponding to 128 isolates

Supplementary 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. The three isolates of the ST399 O25B (ED2020-11-02-e and ED2020-11-02-h) and ST399 O25A (ED2019-09-02-h) are indicated by red and blue arrows, respectively.

Fig. S2. Distribution of the pairwise core genome SNP distances according to whether pairs of isolates are inter- or intra-haplogroup.

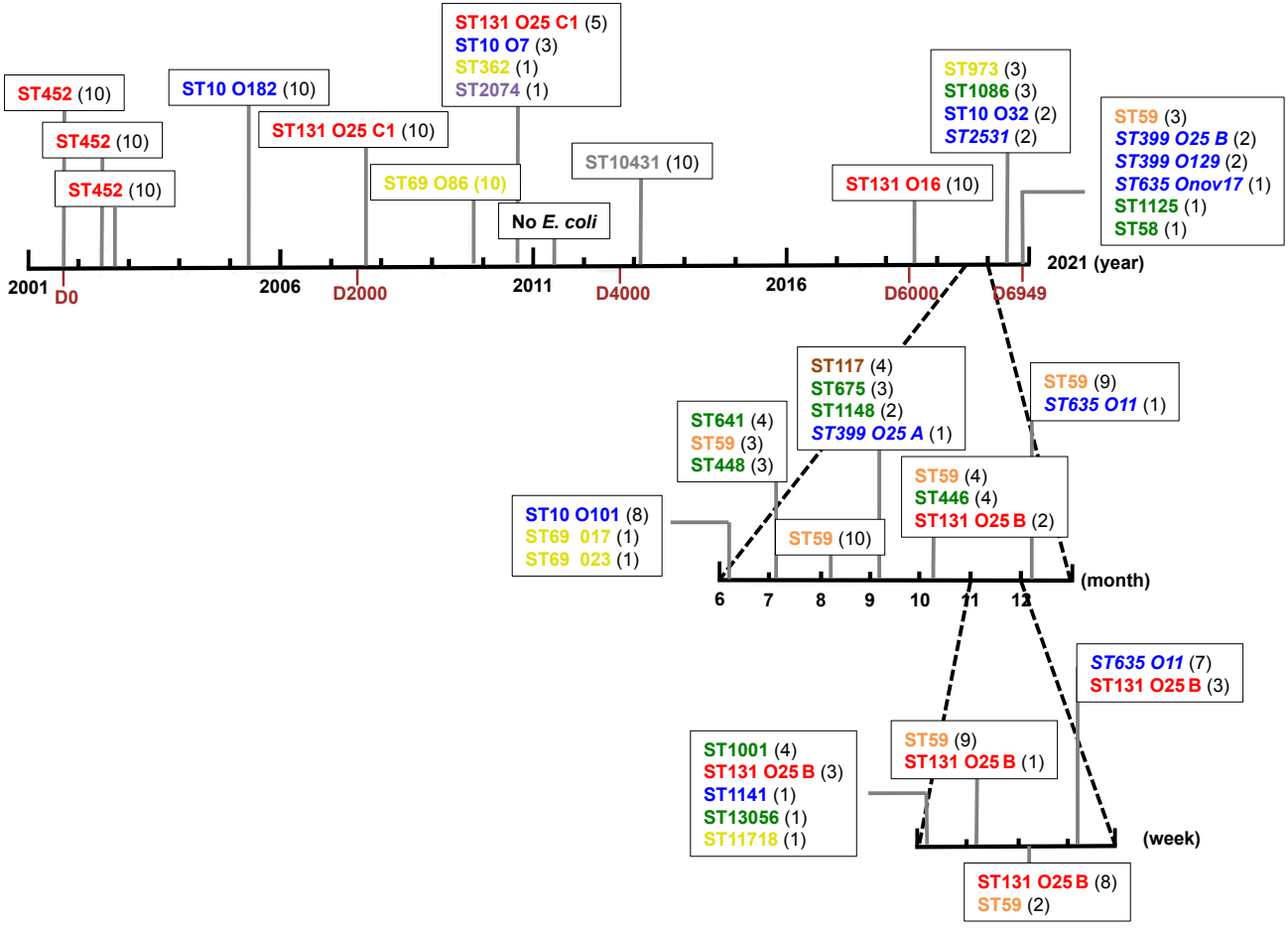
For clarity, the SNP scale is logarithmic with 0.1 for 0. The red dotted line corresponds to the cutoff chosen for the definition of clones. The dots circled in red correspond to the phylogroup A_ ST399/661_O25:H12_fimHND haplogroup isolates differing by 125-130 SNPs.

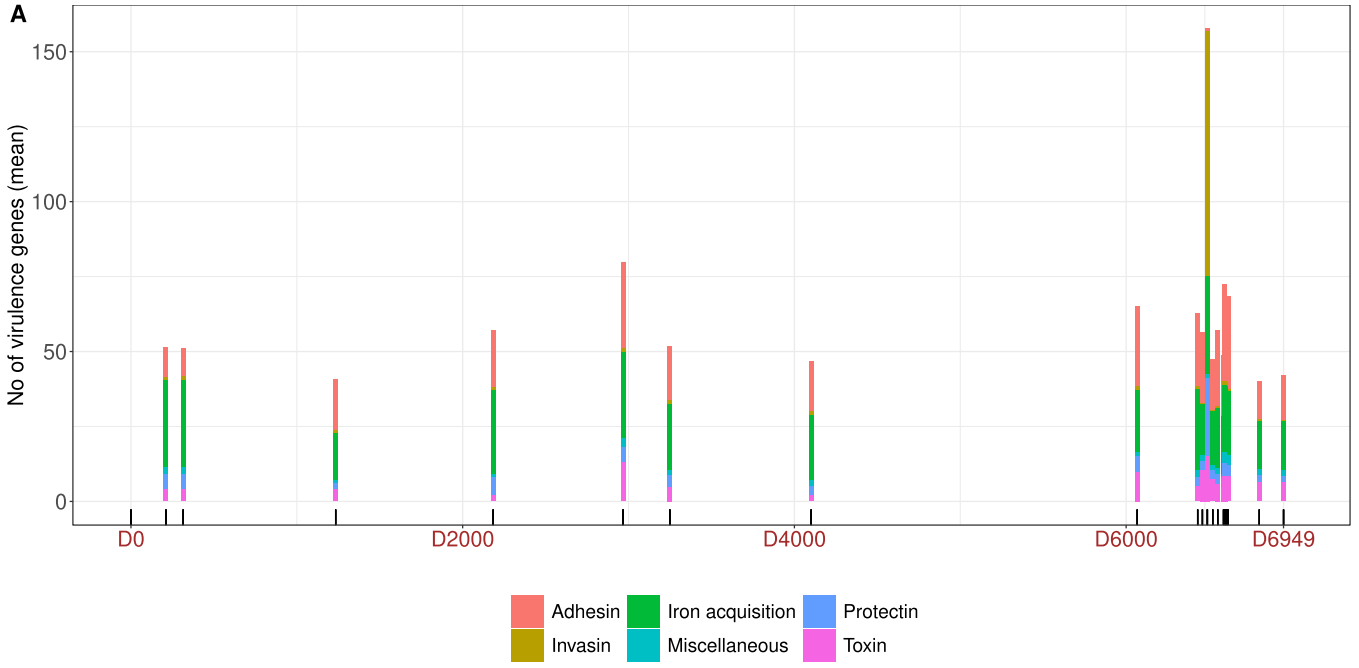
Fig. S3. Mean of the cumulative number of replicons per sample of the 210 *E. coli* ED isolates over time. A, B and C panels correspond to the year, month and week scale of sampling, respectively, as in Fig. 1.

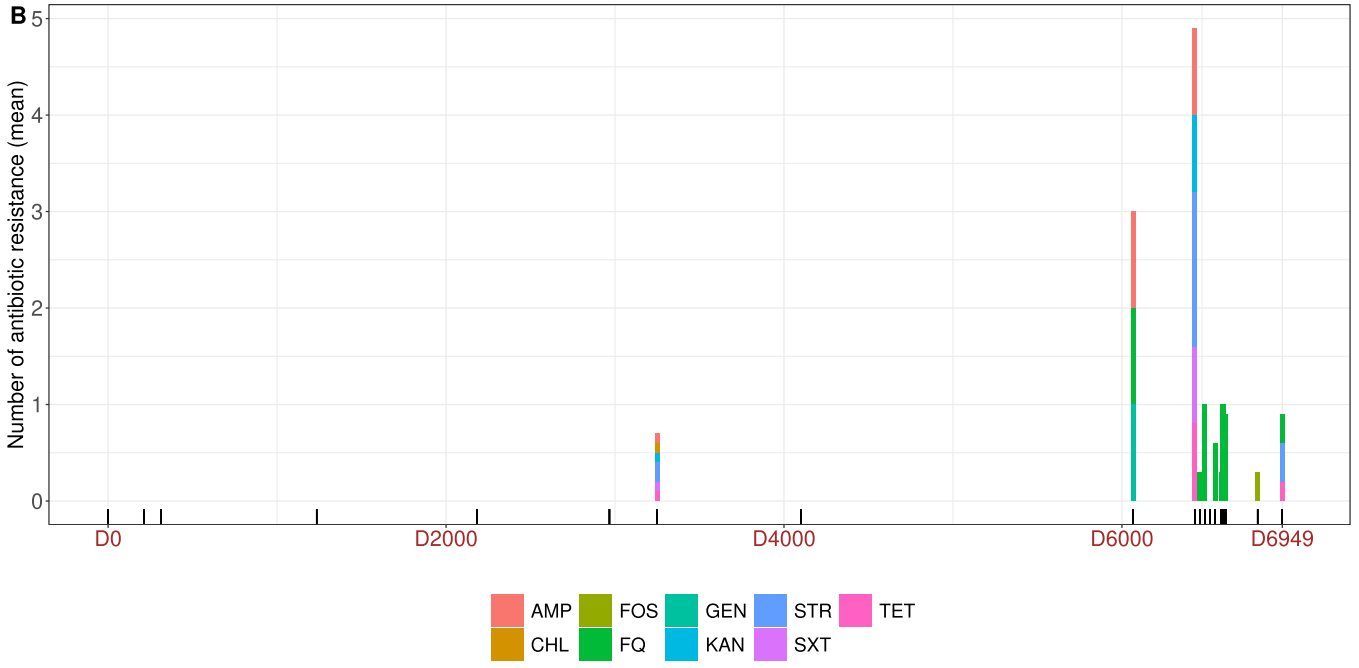
Fig. S4. Distribution of resistance, virulence genes and anti-phage defense systems between ED clones and commensal strains. Mean number of resistant and virulence genes, and total number of defense genes in the 35 ED clones (vertical lines), together with the null distribution when sampling 35 strains at random among the 359 commensals of the control collection (histograms)**.** The ED clones count slightly less resistant genes and less virulence genes, but a similar number of defense systems. The sampling of the control collection was done 50,000 times.

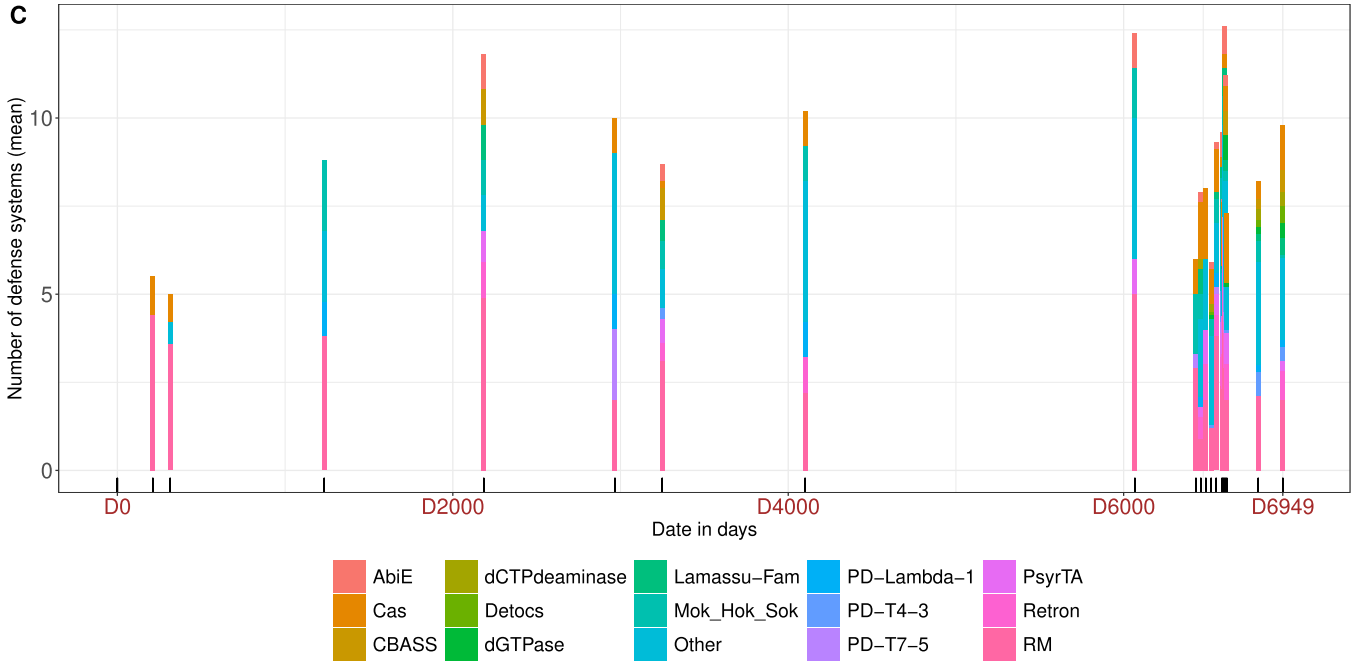
Fig. S5. **Percentage of strains encoding each defense system between the ED clones and the control commensal collection.** Only two systems are significantly more abundant in the ED clones (chi2 test corrected by Bonferroni, significance threshold $p < 0.05$): dGTPase and Detocs.

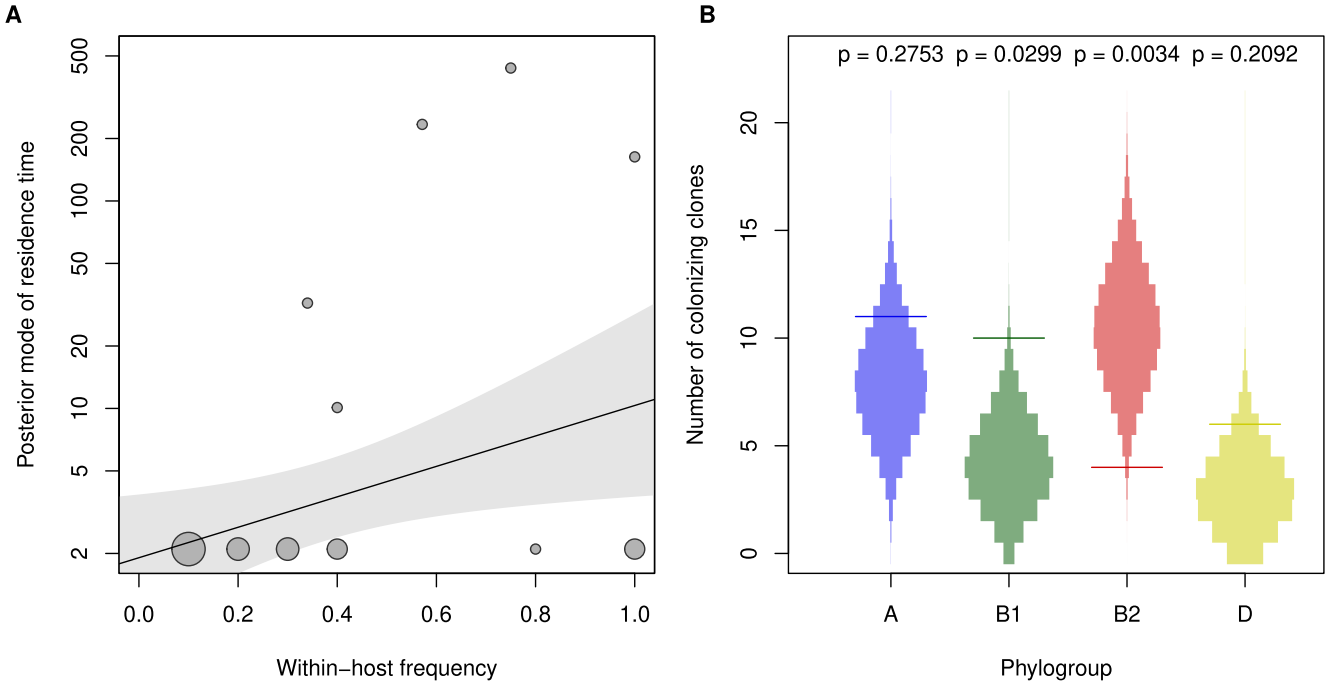
Fig. S6. Physical map of the *mhp* cluster and its genetic environment among phylogroup B2 representative genomes. Representative genomes from phylogroup B2 were selected from the RefSeq dataset for each *mhp* environment based on nucleotide clustering (identity=99%, coverage=95%). The ED1a ST452 genome (GCF_000026305.1) was also included. White arrows represent the *mhp* gene cluster. Genes that are homologous between different genomes are the same colour. The links between the maps represent the identity of the proteins. Insertion sequences are highlighted in red (IS*1*), blue (IS*3*) and green (IS*110*). The combination of serotypes and *fimH* allele was used to assign genomes belonging to ST131 to the clades A (O16:H5/*fimH41*, B (O25B:H4/*fimH161,* O25B:H4/*fimH22*) or C (O25B:H4/*fimH41*).

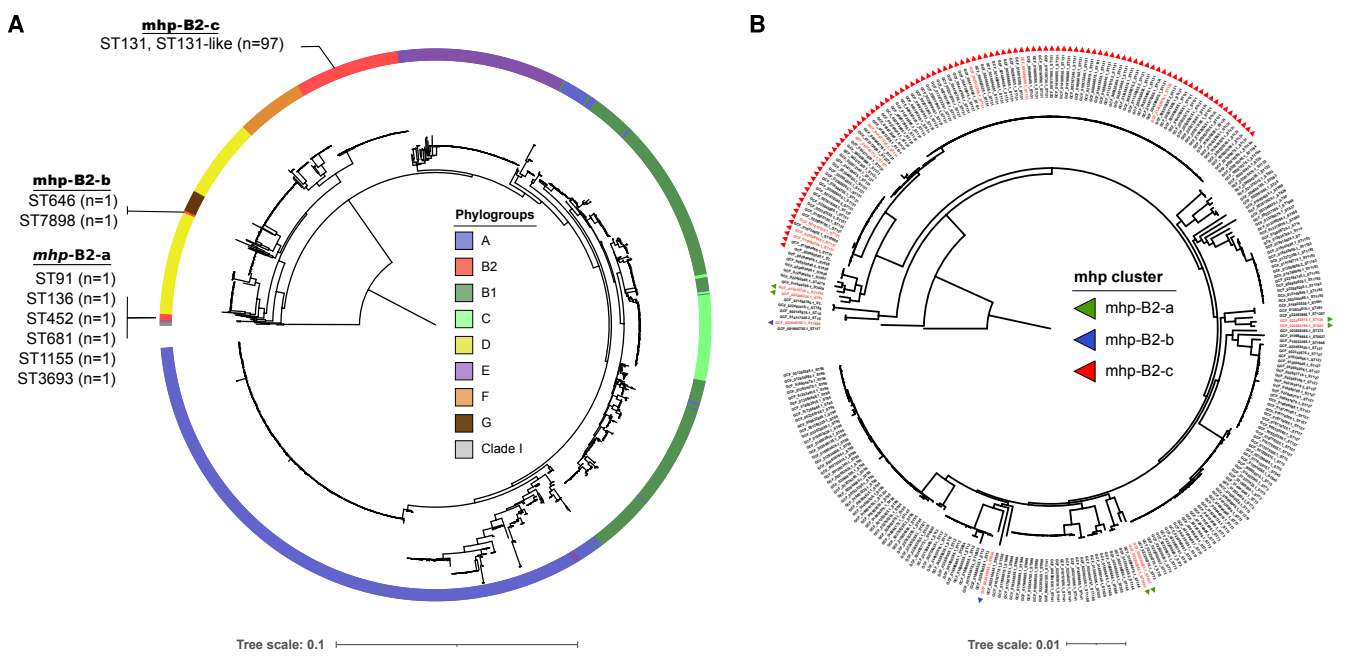

ST452 (10)
ST452 (10)
ST452 (10)
ST10 O182 (10)
ST131 O25 C1 (10)
ST69 O86 (10)
ST131 O25 C1 (5)
ST10 O7 (3)
ST362 (1)
ST2074 (1)
No E. coli
ST10431 (10)
ST131 O16 (10)
ST973 (3)
ST1086 (3)
ST10 O32 (2)
ST2531 (2)
ST59 (3)
ST399 O25 B (2)
ST399 O129 (2)
ST635 Onov17 (1)
ST1125 (1)
ST58 (1)
2001
D0
2006
D2000
2011
D4000
2016
D6000
D6949
2021 (year)
ST117 (4)
ST675 (3)
ST1148 (2)
ST399 O25 A (1)
ST59 (9)
ST635 O11 (1)
ST641 (4)
ST59 (3)
ST448 (3)
ST59 (4)
ST446 (4)
ST131 O25 B (2)
ST10 O101 (8)
ST69 017 (1)
ST69 023 (1)
ST59 (10)
6
7
8
9
10
11
12
(month)
ST635 O11 (7)
ST131 O25 B (3)
ST1001 (4)
ST131 O25 B (3)
ST1141 (1)
ST13056 (1)
ST11718 (1)
ST59 (9)
ST131 O25 B (1)
(week)
ST131 O25 B (8)
ST59 (2)

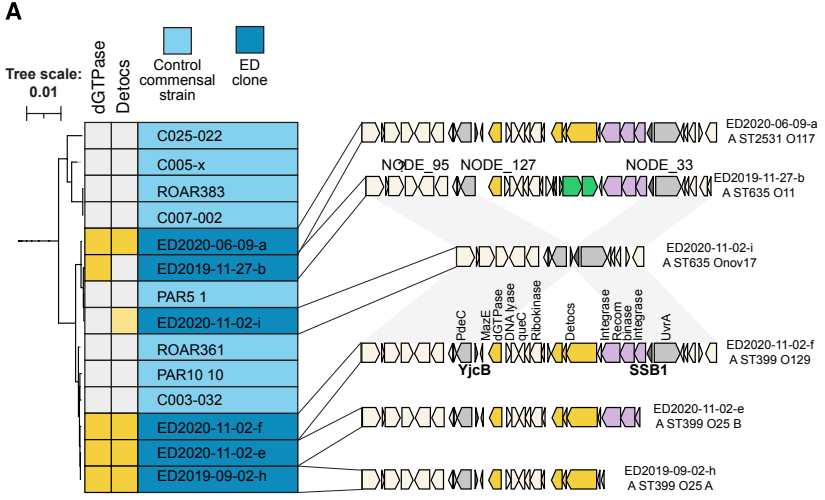

A
No of virulence genes (mean)
150
100
50
0
D0
D2000
D4000
D6000
D6949
Adhesin
Iron acquisition
Protectin
Invasin
Miscellaneous
Toxin

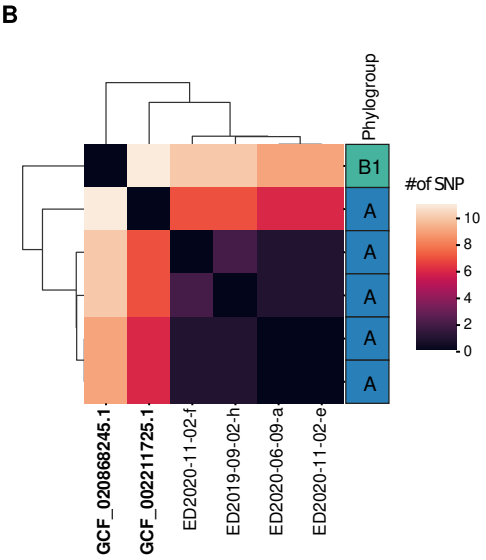

B
Number of antibiotic resistance (mean)
5
4
3
2
1
0
D0
D2000
D4000
D6000
D6949
AMP
FOS
GEN
STR
TET
CHL
FQ
KAN
SXT

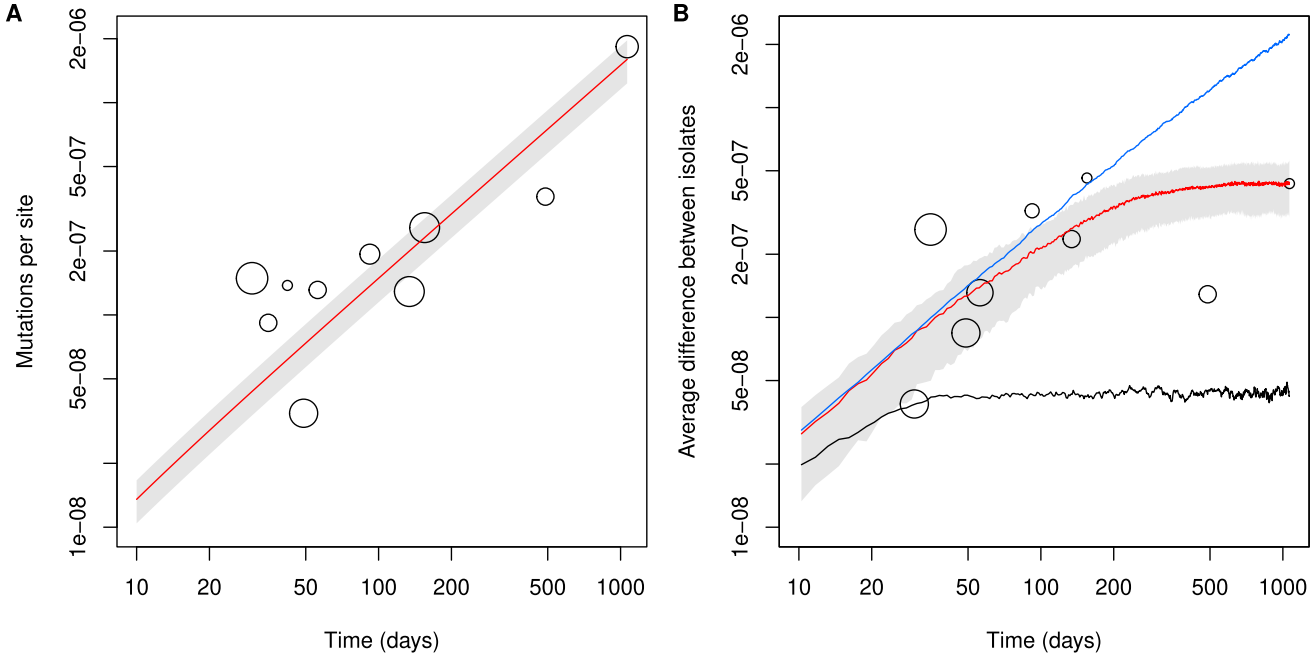

C
Number of defense systems (mean)
10
5
0
D0
D2000
D4000
D6000
D6949
Date in days
AbiE
dCTPdeaminase
Lamassu−Fam
PD−Lambda−1
PsyrTA
Cas
Detocs
Mok_Hok_Sok
PD−T4−3
Retron
CBASS
dGTPase
Other
PD−T7−5
RM


A
Posterior mode of residence time
500
200
100
50
20
10
5
2
0.0
0.2
0.4
0.6
0.8
1.0
Within−host frequency
B
p = 0.2753 p = 0.0299 p = 0.0034 p = 0.2092
Number of colonizing clones
0
5
10
15
20
A
B1
B2
D
Phylogroup


A
mhp-B2-c
ST131, ST131-like (n=97)
mhp-B2-b
ST646 (n=1)
ST7898 (n=1)
mhp-B2-a
ST91 (n=1)
ST136 (n=1)
ST452 (n=1)
ST681 (n=1)
ST1155 (n=1)
ST3693 (n=1)
Phylogroups
A
B2
B1
C
D
E
F
G
Clade I
Tree scale: 0.1
B
mhp cluster
mhp-B2-a
mhp-B2-b
mhp-B2-c
Tree scale: 0.01


A
Tree scale:
0.01
dGTPase
Detocs
Control commensal strain
ED clone
C025-022
C005-x
ROAR383
C007-002
ED2020-06-09-a
ED2019-11-27-b
PAR5 1
ED2020-11-02-i
ROAR361
PAR10 10
C003-032
ED2020-11-02-f
ED2020-11-02-e
ED2019-09-02-h
NODE_95
NODE_127
NODE_33
PdeC
MazE
dGTPase
DNA lyase
queC
Ribokinase
Detocs
Integrase
Recom-binase
Integrase
UvrA
YjcB
SSB1
ED2020-06-09-a
A ST2531 O117
ED2019-11-27-b
A ST635 O11
ED2020-11-02-i
A ST635 Onov17
ED2020-11-02-f
A ST399 O129
ED2020-11-02-e
A ST399 O25 B
ED2019-09-02-h
A ST399 O25 A


B
Phylogroup
B1
A
A
A
A
A
#of SNP
10
8
6
4
2
0
GCF_020868245.1
GCF_002211725.1
ED2020-11-02-f
ED2019-09-02-h
ED2020-06-09-a
ED2020-11-02-e


A
Mutations per site
1e−08
5e−08
2e−07
5e−07
2e−06
10
20
50
100
200
500
1000
Time (days)
B
Average difference between isolates
1e−08
5e−08
2e−07
5e−07
2e−06
10
20
50
100
200
500
1000
Time (days)