Residence-colonization trade-off and niche differentiation enable the coexistence of *Escherichia coli* phylogroups in healthy humans

Thibaut Morel-Journel[a,1], Sonja Lehtinen[b], Olivier Cotto[c], Rafika Amia[d], Sara Dion[d], Clarisse Figueroa[d], Jonathan Martinson[e], Pascal Ralaimazava[f], Olivier Clermont[d], Xavier Duval[d,f], Forough L. Nowrouzian[g], Seth Walk[h], Erick Denamur[d,i], and François Blanquart[j]

[a]Université Sorbonne Paris Nord, Université Paris Cité, INSERM, IAME, Bobigny, France
[b]Department of Environmental System Science, Institute for Integrative Biology, ETH Zürich, Zürich, Switzerland
[c]Plant Health Institute of Montpellier, INRAE, Univ Montpellier, CIRAD, Institut Agro, IRD, Montpellier, France
[d]Université Paris Cité, Université Sorbonne Paris Nord, INSERM, IAME, Paris, France
[e]Innovative Genomics Institute, University of California, Berkeley, CA, USA
[f]AP-HP, Hôpital Bichat, Centre d'Investigation Clinique, INSERM CIC 1425, Université Paris Cité, France
[g]Department of Infectious Diseases, Institute of Biomedicine, University of Gothenburg, Guldhedsgatan 10, S-413 46, Gothenburg, Sweden
[h]Department of Microbiology and Immunology, Montana State University, Bozeman, MT, USA
[i]AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, Paris, France
[j]CIRB, Collège de France, Université PSL, CNRS, INSERM, Paris, France

Keywords

Coexistence, *Escherichia coli*, Life-history trade-off, Microbiota

Abstract

Despite extensive literature on the pathogenicity and virulence of the opportunistic pathogen *Escherichia coli*, much less is known about its ecological and evolutionary dynamics as a commensal in healthy hosts. Based on two detailed longitudinal datasets on the gut microbiota of healthy adult individuals followed over months to years in France and in the USA, we identified a robust trade-off between the ability to establish in a new host (colonization) and the ability to remain in the host (residence). Major *E. coli* lineages (phylogroups) exhibited similar fitness but a diversity of strategies, from strong colonizers residing for a few days in the gut, to poor colonizers residing for years. Strains with the largest number of extra-intestinal virulence associated genes and highest pathogenicity resided for longest in hosts. Moreover, the residence time of a strain was reduced more strongly when it competed with other strains of the same phylogroup than of different phylogroups, suggesting niche differentiation between them. To investigate the consequences of the trade-off and niche differentiation for coexistence between strains, we developed a discrete-state Markov model describing the dynamics of *E. coli* in a population of hosts. We found that the trade-off and niche differentiation acted together as equalizing and stabilizing mechanisms enabling the coexistence of phylogroups over extended periods of time. Our model predicted that a reduction in transmission (e.g. better hygiene) would not alter the balance between phylogroups, while disturbance of the microbiome (e.g. antibiotics) would hinder residents strains such as those of the extra-intestinal pathogenic phylogroup B2.3.

Introduction

Understanding how bacteria diversify under the combined action of mutation and recombination is a long-standing issue in microbiology and evolutionary biology [1, 2]. Multiple mechanisms have been proposed to explain the observed diversity and the coexistence of bacterial strains that may exhibit fitness differences. Trade-offs between life-history traits can enable species to diversify into distinct strategies [3, 4]. Differentiation between ecological niches among bacteria can also favor diversity by allowing distinct strains to exploit the heterogeneity between hosts [5, 6]. For instance, the combination of a serotype-specific and non-specific host immune responses can generate host differences allowing coexistence among *Streptococcus pneumoniae* serotypes [7]. Other mechanisms not involving selection, such as a recombination rate negatively correlated with genetic distance, can lead to genetically distinct groups as well [1, 8].

The question of coexistence is especially critical for species such as *Escherichia coli*, which is not only a commensal bacteria of the human gut [9], but also an opportunistic pathogen. *E. coli* is responsible for various forms of diarrhea [10, 11] and extra-intestinal infections, including urinary tract [12, 13] and bloodstream infections [14, 15], causing around one million deaths worldwide each year [16–18]. The species is highly diverse and structured [19, 20], notably around lineages called 'phylogroups' [21, 22], which are largely monophyletic [9, 23] and differ in their virulence and antibiotic resistance profiles [24, 25]. Commensal and extraintestinal pathogenic *E. coli* are not strictly phylogenetically separated [26] and the transition to infection is the result of complex and hardly predictable interactions between several genetic, host and environmental factors [27]. Infections remain very rare events among the species, so that genes associated with pathogenicity (the probability to cause an infection) and virulence (the severity of infection) may evolve mainly as a result of their role in commensalism [28, 29]. Characterizing the dynamics of commensal *E. coli*, as well as their association with pathogenicity and virulence, is essential for understanding how the phylogenetic structure of *E. coli* emerged and for predicting the future evolution of the species.

This calls for precise characterization of the life-history traits of *E. coli*, which define the epidemiological dynamics in the human population – the rates at which strains colonize new hosts are cleared from them. These dynamics can be quantified through longitudinal follow-up of *E. coli* in healthy subjects, which unfortunately remains rare. Studies rather focus on diseased individuals [e.g. 30, 31], for example to assess the impact of antibiotic use on the microbiota [32, 33]. The initial colonization of infants during their first months of life has been studied [34–37], but longitudinal studies on healthy adults include a limited number of subjects [38–42].

The scarcity of longitudinal studies on healthy individuals can be explained by the complexity of their implementation. In addition to the need for a large number of subjects and of samples by subject, a single host can harbor several strains at the same time [43], multiplying the amount of work required to characterize each sample. In this study, we resolved this problem by using two complementary and remarkably detailed longitudinal datasets to characterize the dynamics of *E. coli* in healthy adults of high-income countries. The first one, previously analyzed by Martinson et al. [44], includes eight subjects followed over months to years in the USA with an extremely thorough characterization of the different clones present at the same time in each host. The second dataset, original to this study, includes 46 subjects followed over four to five months in France. Based on these datasets, we aimed to (i) quantitatively characterize the life-history traits of *E. coli* phylogroups, (ii) correlate these commensal traits with those associated with pathogenicity and virulence in the species and (iii) derive the consequences of our results for the epidemiological dynamics and potential coexistence of these phylogroups.

Results

Survival analyses

The first dataset we used to study the turnover dynamics of commensal *E. coli* (called thereafter the 'French dataset') included 415 clones sampled from 46 hosts over periods of 111 to 135 days. The second one (called thereafter the 'USA dataset') included 117 clones sampled from 8 hosts over periods of 245 to 849 days. We used the Clermont method [21, 22] to assign the clones of both datasets to phylogroups and additional PCR-based DNA fingerprinting techniques to differentiate clones of a same phylogroup within a single host (see Material and Methods).

We used survival models, specifically parametric accelerated failure time (AFT) models, to assess the impact of various factors on the rate at which clones colonized new hosts and their residence time once in their host. We performed one analysis per dataset (French or USA) and per response variable (colonization rate or residence time). For each analysis, we used the Akaike's information criterion [AIC, 45] to select the combination of explanatory factors best explaining the data (see Supporting information). The explanatory factors potentially included in the best model were the following: the phylogroup of the clone considered, its host, the total number of other clones in the host, the number of other clones of the same phylogroup in the host and the total cell density in the host. We included the cell density of the clone considered in the analysis of residence time in the USA data only, as no precise estimate was available for the French data owing to the smaller number of colonies typed per sample.

Determinants of the residence time

For both datasets, AFT models showed differences in clone residence times according to host and phylogroup (see Table S2 for the French dataset and Table S4 for the USA dataset). They also provided support for greater niche overlap between clones of the same phylogroup than between those of different phylogroups. Indeed, the residence time of a clone was significantly reduced by the presence of other clones in the host, but even more so if these other clones were of the same phylogroup (Table 1). According to the analysis of the French dataset, the residence time was divided by 1.45 (95%CI = $[1.219, 1.717]$) for each additional clone of a different phylogroup, and by 2.09 (95%CI = $[1.594, 2.751]$) for each additional clone of the same phylogroup. According to the analysis of the USA dataset, the residence time was divided by 1.68 (95%CI = $[1.266, 2.358]$) for each additional clone of different phylogroup, and by 5.09 (95%CI = $[1.800, 14.423]$) for each additional clone of the same phylogroup. The analysis of the USA dataset also showed that the cell density of a clone increased its residence time (Table 1), which was multiplied by 1.32 (95%CI = $[1.002, 1.499]$) when the clone density was multiplied by 10. Neither of

Table 1: Z-score and p-values of the coefficients corresponding to the continuous explanatory variables in the AFT models of residence and colonization selected for the analyses of the French and USA datasets. Positive value of z-score indicate longer residence times/lower colonization rates as the factor increases. See Supporting information for the tables including qualitative factors, i.e. host and phylogroup.

Dataset	Response variable	Explanatory factor	z-score	p-value
French	Residence time	Total number of clones	-4.218	< 0.0001
French	Residence time	Number of clones of the same phylogroup	-2.654	0.0079
USA	Residence time	Total number of clones	-3.444	0.0006
USA	Residence time	Number of clones of the same phylogroup	-2.135	0.0328
USA	Residence time	Log_{10} of cell density	1.983	0.0473
French	Colonization rate	Total number of clones	-9.493	< 0.0001
USA	Colonization rate	Total number of clones	-9.775	< 0.0001

the two AFT models selected based on AIC included total cell density as an explanatory factor.

To confirm the robustness of these results to the statistical method used, we performed additional analyses using semi-parametric Cox models with the same explanatory factors as those used for the AFT models, which yielded consistent results (see Supporting information).

Determinants of the colonization rate

While we assumed that residence time was impacted by within-host dynamics alone, we considered colonization to be the result of inter-host transmission. Therefore, we expected the colonization rate of a phylogroup to be affected by its own frequency in the host population as a whole. To account for this, we corrected the colonization rates observed by the frequencies of phylogroups in their respective countries, based on cross-sectional external data (see Supporting information). We then performed the survival analyses using these corrected data.

As for residence times, AFT models showed differences in colonization rates according to host and phylogroup for both datasets (see Table S7 for the French dataset and Table S9 for the USA dataset). The two models also showed that the colonization rate was increased by the presence of other clones in the host (Table 1). The colonization rate was predicted to be multiplied by 1.76 (95%CI $= [1.562, 1.970]$) for each additional clone according to the analysis of the French dataset, and by 2.89 (95%CI $= [2.342, 3.589]$) for each additional clone according to the analysis of the USA dataset. However, the impact of other clones already residing in the host on the colonization rate was phylogroup-independent. Indeed, neither of the two AFT models selected based on AIC included the number of clones of the same phylogroup. They also did not select cell density in the host as an explanatory factor.

Again, we performed additional analyses using semi-parametric Cox models with the same explanatory factors as those used for the AFT models, which confirmed the results of the AFT models (see Supporting information).

Trade-off between colonization and residence

To investigate the relationship between residence and colonization at the phylogroup level, we computed the estimates and standard errors of colonization rate and residence time predicted by the AFT models for each phylogroup (Fig 1) and quantified the association between them. The two life-history traits were negatively associated, both in France (Spearman's $\rho = -0.615$, parametric bootstraping 95%CI $= [-0.769, -0.238]$, $p = 0.0016$) and in the USA (Spearman's $\rho = -0.783$, parametric bootstraping 95%CI $= [-0.917, -0.599]$, $p < 0.0001$). This suggested a negative trade-off between the colonization and residence abilities at the phylogroup level. In other words, phylogroups were distributed along an axis ranging from those with a fast turnover (strong colonizers with short residence times) to those with a slow turnover (poor colonizers with long residence times). The arrangement of phylogroups along this trade-off was broadly conserved in France and the USA, with slow-turnover phylogroups including B2.3, D and F; intermediate phylogroups including A.2 and B1, and fast-turnover phylogroups including B2.1 (and B2.2 in the USA only, G in France only).

To further assess the robustness of the trade-off, we also analyzed colonization using colonization rates corrected by the frequencies of phylogroup in the datasets themselves rather than external data (see Supporting information). Using estimates from these models, we found similar negative correlations with the residence time, for the analyses of the French dataset (Spearman's $\rho = -0.769$, parametric bootstraping 95%CI $= [-0.832, -0.336]$, $p = 0.0002$) and the USA dataset (Spearman's $\rho = -0.733$, parametric bootstraping 95%CI $= [-0.900, -0.583]$, $p < 0.0001$).

Finally, we investigated whether the trade-off we identified in adults was also observed during the initial gut colonization of infants (see Supporting information). We analyzed an additional longitudinal

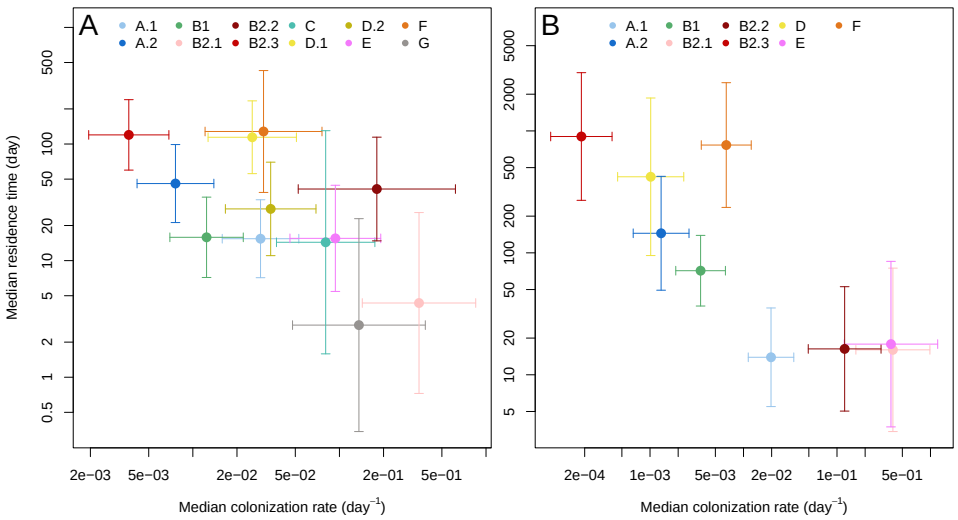

Figure 1: Median residence time as a function of the median colonization rate for each phylogroup appearing in the French (A) and the USA (B) datasets, for the average host and average values of the quantitative explanatory factors in their respective datasets. Estimates (dots) are represented with their 95% confidence interval (bars).

dataset including infants sampled over a year after birth, previously analyzed by Ostblom et al. [36]. Results indicated that this trade-off was also visible in this specific context.

Correlation between residence, pathogenicity and virulence-associated genes

The trade-off between colonization and residence has implications for the evolution of bacterial virulence, as the residents are also the most pathogenic bacteria, and tend to carry more virulence-associated genes (VAG) [46, 47]. We estimated pathogenicity (the odds ratio for the association between bloodstream infections and phylogroup) and the number of VAG carried for each phylogroup, based on a dataset previously analyzed by Marin et al. [48] and Burgaya et al. [49]. The estimated pathogenicity of phylogroups was positively correlated with residence for both datasets (French dataset: Spearman's $\rho = 0.518$, parametric bootstraping 95%CI $= [0.105, 0.741]$, $p = 0.0131$; USA dataset: Spearman's $\rho = 0.745$, parametric bootstraping 95%CI $= [0.217, 0.833]$, $p = 0.0035$)(Fig 2A, 2B). The total number of VAG carried was positively correlated with residence for the French data (Spearman's $\rho = 0.273$, parametric bootstraping 95%CI $= [0.049, 0.531]$, $p = 0.0221$) and the USA data suggested the same trend (Spearman's $\rho = 0.383$, parametric bootstraping 95%CI $= [-0.017, 0.533]$, $p = 0.0527$)(Fig 2C). Differences were observed in the strength of the correlation when considering genes associated with specific functions: those linked with protectins were the most strongly correlated with residence, while those linked with toxins were not. Genes linked with iron acquisition ($p = 0.0473$) and to adhesins ($p = 0.0178$) were significantly correlated with residence, for the French and USA datasets respectively.

Model of bacterial epidemiological dynamics

Following data analysis, we derived the implications of our results for the epidemiological and evolutionary dynamics of *E. coli*. We developed a continuous time, discrete-state Markov chain model describing the dynamics of clones in a host population and integrating our findings on colonization rates and residence times. In this model, discrete states were sets of the phylogroups of *E. coli* clones potentially present in the hosts. Phylogroups were numbered from 1 to G and states were numbered in increasing order of the number of clones in the host. For instance, an 'empty' host (without any clone) was in state $S_1 = \{\emptyset\}$,

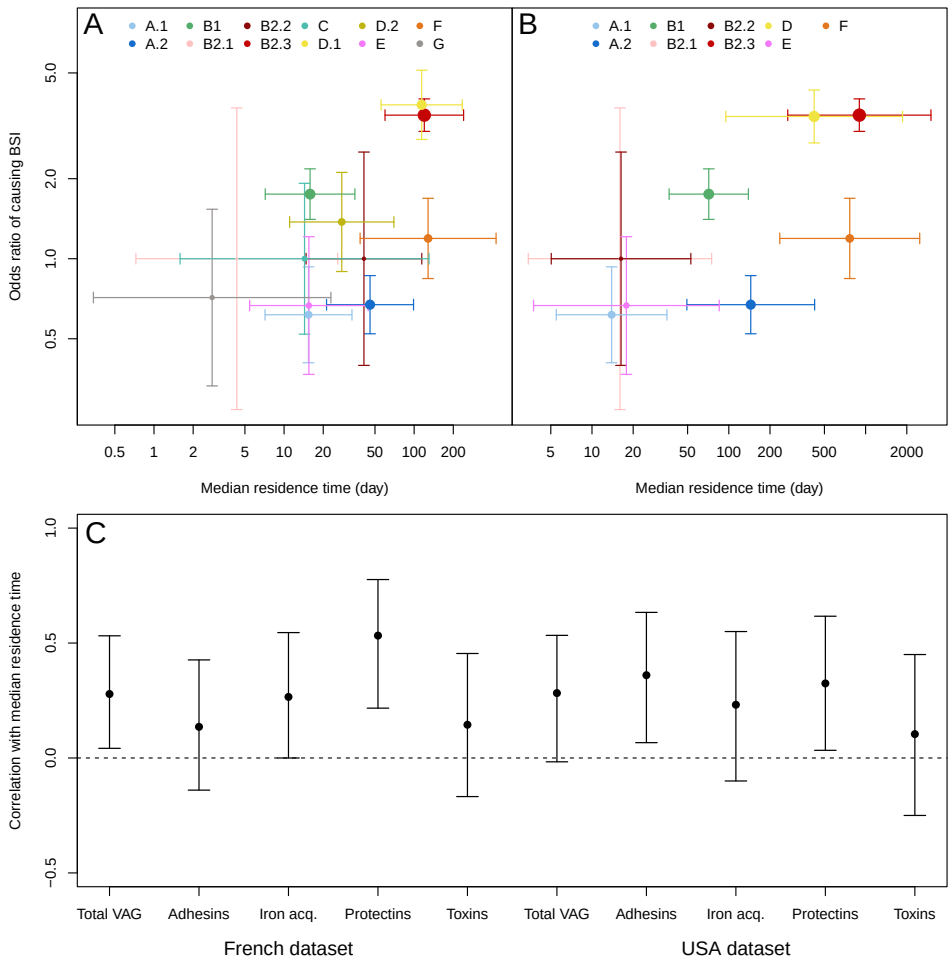

Figure 2: A and B: odds ratio of the association with bloodstream infections (BSI) as a function of median residence time for each phylogroup appearing in the French (A) and the USA (B) datasets. Estimates (dots) are represented with their 95% confidence interval (bars), with the size of the dots proportional to the number of samples of a given phylogroup used to estimate the odds ratios of causing BSI. C: correlation rate between the median residence time of each phylogroup estimated for the two datasets and its number of VAG (total or associated with adhesins, iron acquisition systems, protectins or toxins). Estimates (dots) are represented with their parametric bootstrap 95% confidence interval (bars). Correlations whose confidence interval does not include 0 are considered significant.

a host colonized by a single clone of phylogroup 1 in state $S_2 = \{1\}$, a host colonized by a single clone of phylogroup 2 in state $S_3 = \{2\}$ and a host colonized by two clones of group 1 in state $S_{G+2} = \{1, 1\}$. The variable of the model was the probability $P_i(t)$ that a host was in state S_i at time t, which could also be interpreted as the proportion of an infinite population of hosts in state S_i at this time.

Transitions between states depended on clearance rates μ_{k,n,n_k} and colonization rates $\lambda_{k,n}(t)$:

$$\begin{aligned} \mu_{k,n,n_k} &= 1/e^{a_k + b(n-1) + b_s(n_k - 1)} \\ \lambda_{k,n}(t) &= F_k(t)/e^{c_k + dn} \end{aligned} \quad . \tag{1}$$

As inferred from the data, clearance was affected by three parameters: a_k was phylogroup-specific, b represented the impact of the number of other clones in the host $n - 1$ and b_s represented the impact of the number of other clones of the same phylogroup $n_k - 1$. In the French and USA datasets, we inferred negative values for b – clearance was faster the more clones were present in the host – and for b_s – clearance was faster the more other clones present belonged to the same phylogroup. The colonization

rate of a clone was affected by the frequency of the phylogroup in the host population $F_k(t)$ and two parameters: c_k was phylogroup-specific and d represented the impact of the number of other clones in the host $n-1$. In the French and USA datasets, we inferred negative values for d, meaning that colonization was easier when more clones were present in the host.

Each phylogroup was characterized by an intrinsic colonization-to-clearance ratio $R_k = e^{a_k - c_k}$, which was also its basic reproduction number. Indeed, when a single clone of phylogroup k was introduced in a large population of H empty hosts, the frequency of this clone was $F_k(t) = 1/H$, the number of hosts to colonize was approximately H and the expected residence time of the clone was $1/\mu_{k,1,1}$. Thus, the expected number of hosts colonized after this single introduction was:

$$H\lambda_{k,0}(t) \times \frac{1}{\mu_{k,1,1}} = H\frac{(1/H)\, e^{-(c_k+0\times d)}}{e^{-(a_k+0\times b_s+0\times b_o)}} = e^{a_k-c_k} = R_k. \tag{2}$$

When competition between clones was the same irrespective of phylogroup ($b_s = 0$), the colonization-to-clearance ratio also predicted which phylogroup could invade others (see Supporting information). Indeed, the differences in colonization-to-clearance ratio R_k defined a 'hierarchy' between phylogroups, so that those with higher ratios could invade those with lower ratios. The existence of a trade-off then corresponded to several phylogroups with the same colonization-to-clearance ratio R_k but different strategies, e.g. from a fast turnover (high values of a_k and c_k) to a slow turnover (low values of a_k and c_k).

We explored with this mathematical model how this hierarchy between phylogroups was affected by non phylogroup-specific public hygiene measures (see Supporting information). On the one hand, the hierarchy was not affected by a measure resulting in a 'multiplicative' change, i.e. by multiplying the colonization or clearance rate by a constant factor. On the other hand, a measure adding a new source of clearance could alter the hierarchy of phylogroups by disproportionately affecting the phylogroup with a slower turnover. These changes in hierarchy could translate into variations in phylogroup frequencies. For instance, increasing the strength of this source of clearance eventually lead to the dominance of a fast-turnover phylogroup over a slow-turnover one (Fig 3).

Conditions for the coexistence of multiple phylogroups

We next assessed the conditions in which our model enabled coexistence between two phylogroups, depending on their respective colonization-to-clearance ratios. We quantified coexistence with two measures. Within-host coexistence $E^W(t) \in [0,1]$ corresponded to the proportions of hosts carrying clones of at least two phylogroups. Inter-host coexistence $E^I(t) \in [0, \ln(G)]$ was a Shannon index on the relative frequencies of the phylogroups in the host population:

$$E^I(t) = -\sum_{k=1}^{G} f_k^h(t) \ln\left(f_k^h(t)\right), \quad f_k^h(t) = \frac{F_k^h(t)}{\sum_{k=1}^{G} F_k^h(t)}. \tag{3}$$

We considered a simple case of two phylogroups i and j, characterized by an average colonization-to-clearance ratio $\overline{R}_{i,j} = (R_i + R_j)/2$ and a ratio difference $D_{i,j} = R_i - R_j$. We performed discrete-time simulations of our model with a time-step of one day until an asymptotic state was reached. We assumed that this state to be reached when state probabilities essentially no longer varied, i.e. at time t^* such that $\sum_{i=1}^{N} |P_i(t^*) - P_i(t^*-1)| < 10^{-12}$. We computed the asymptotic within-host coexistence E^{W*} and inter-host coexistence E^{I*} at the steady state.

The results showed that only structurally unstable coexistence was possible if there was no additional competition between clones of the same phylogroup ($b_s = 0$, see Eq. 1), which was in line with the mathematical analysis of the model (see Supporting information). In that case, coexistence was maintained

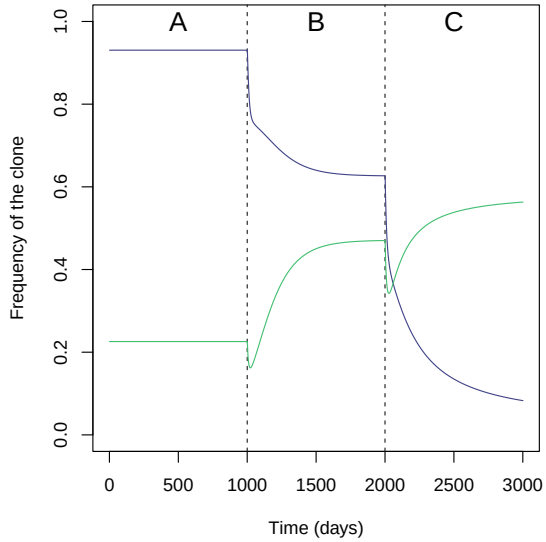

Figure 3: Variation in the frequencies of phylogroups x (purple) and y (green) over 3000 simulated days of the model with increasing levels of additive clearance (θ) , with $R_x = 3.5$, $R_y = 3$, $b = 0$, $b_s = -0.1$ and $d = 0$. Three time-periods of 1000 days each (separated by dashed lines) are defined: no additive clearance ($\theta = 0$) for $t \in [0, 1000[$ (A), $\theta = 0.1$ for $t \in [1000, 2000[$ (B) and $\theta = 0.2$ for $t \in [2000, 3000[$ (C). Phylogroup x has a slower turnover than y. the phylogroup hierarchy is reversed for a high enough value of θ.

only if the two phylogroups had exactly the same colonization-to-clearance ratio, i.e. if $D_{i,j} = 0$. Stable coexistence was possible when there was additional competition between clones of the same phylogroup ($b_s < 0$). As b_s became more negative (greater niche differentiation), the range of values of $D_{i,j}$ for which coexistence was stable widened, but within-host coexistence also declined (Fig 4A). Indeed, more negative values of b_s lead to higher average clearance rates overall, thus decreasing overall the number of clones per host, and the value of E^{W*}. Inter-host coexistence was not impacted in the same way and E^{I*} remained high as b_s became more negative (Fig 4B).

The simulation results presented in Fig 4 correspond to $\overline{R}_{i,j} = 2$, $b = 0$ and $d = 0$, but similar results were obtained for other parameter values. Higher values of $\overline{R}_{i,j}$ (the average colonization-to-extinction ratio), more positive values of b (lower clearance rate as the number of clones in the host increases) and more negative values of d (higher colonization rate as the number of clones in the host increases) all resulted in a higher overall number of clones per host, and therefore wider ranges of $D_{i,j}$ for which coexistence was stable for a given value of b_s. Lower values of $\overline{R}_{i,j}$, more negative values of b or more positive values of d had the opposite effect.

Discussion

By analyzing longitudinal datasets on *E. coli* gut colonization dynamics in healthy adults from high-income countries, we identified two mechanisms explaining the stable coexistence of *E. coli* phylogroups. The first one is the trade-off between the colonization and residence abilities at the phylogroup level. Our results indicate that the 'resident' or 'transient' nature of *E. coli* clones, which has been observed for over 100 years [39, 50], is a property partly explained by phylogroups and inversely correlated with the ability to colonize hosts. This trade-off is equalizing *sensu* Chesson [51], meaning that it reduces differences in fitness between the phylogroups. In our model of colonization dynamics with human-to-human transmission, fitness is mathematically equal to the colonization-to-clearance ratio R_k. Therefore, the coexistence of multiple phylogroups in a population at equilibrium implies that these phylogroups

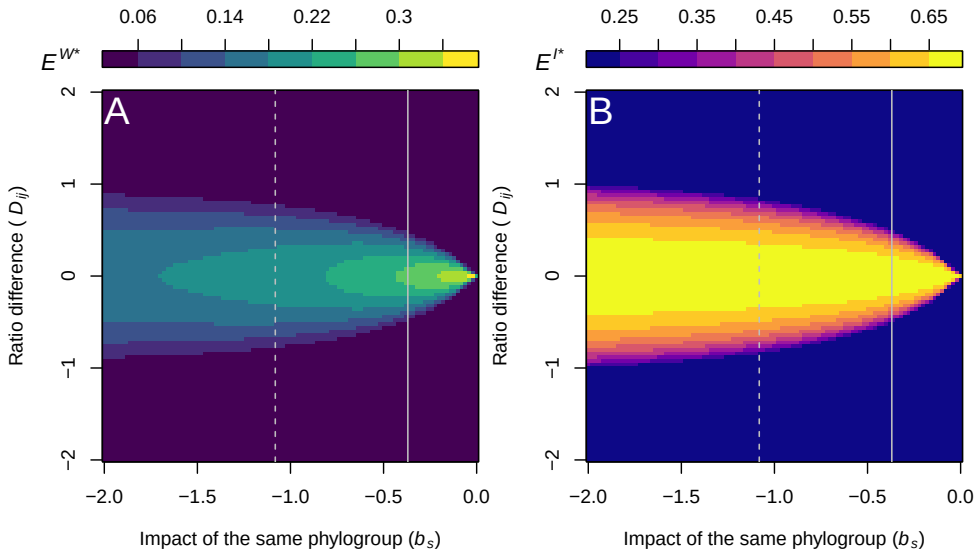

Figure 4: Asymptotic within-host coexistence E^{W*} (A) and inter-host coexistence E^{I*} (B) as a function of the ratio difference $D_{i,j}$ and the strength of clones of the same phylogroup on clearance b_s. Vertical lines correspond to the coefficients characterizing the impact of other clones of the same phylogroup from the best AFT models for the French (solid) and USA datasets (dashed). For these simulations, $\overline{R}_{i,j} = 2$, $b = 0$, $d = 0$, $E^{W*} \in\ ]0, 1]$ and $E^{I*} \in]0, \ln(2)]$.

compensate shorter residence times with a higher colonization rate (or vice versa) to achieve similar fitness. Indeed, clones with lower fitness would not rise in frequency, while clones with larger fitness would invade and replace all other clones. This trade-off was also evidenced in infants followed during their first year of life, who have much simpler microbiome [52], suggesting that it does not depend on the presence of other species.

The second mechanism is niche differentiation, which acts as a stabilizing mechanism that can explain long-term coexistence [51]. In both datasets, residence is shorter when a clone is in a host together with another clone of the same phylogroup. This indicates stronger competition within phylogroups than between phylogroups, which could be explained by a greater overlap of ecological niches between clones that are phylogenetically closer. Interestingly, colonization does not exhibit the same dependence to the number of clones of the same phylogroup. This suggests that the niche overlap observed only affects clones after the colonization of a new host.

Our analyses also highlighted substantial heterogeneity between hosts in bacterial dynamics, consistent with the absence of 'core resident microbiome' initially reported by Martinson et al. [44] using the USA data. Clone turnover rates varied not only between hosts, but also over time within the same host. Both clearance and colonization were more frequent in hosts with a greater number of different clones. This was particularly apparent in the USA dataset, in which hosts experienced periods of stability with few clones, alternating with rapid successions of colonization and clearance events, during which the number of clones present temporarily increased. These periods of instability could be triggered by temporal variations in the suitability of the host for colonization or residence. They could also be triggered by the temporary destabilization of the microbial community due to the colonization of several clones in a short time span. These explanations are not exclusive and could both contribute to temporal variability. These transient periods of instability have also been observed over decades in a single host, with long periods of residence of clones, mainly of phylogroup B2, interrupted by rapid successions of clones, mainly of phylogroup B1 or A [42].

Although this study focuses on commensal *E. coli*, our results also bear implications for the evolution of pathogenicity. Previous studies have highlighted phylogroup B2, massively represented in our data

by B2.3 (see Fig S1) as both resident and virulent [29, 53, 54]. We present a more general positive correlation between the residence time of a phylogroup, and both its pathogenicity and its number of virulence-associated genes. These results are consistent with the hypothesis that the virulence of *E. coli* as an opportunistic pathogen results from traits linked with its residence ability as a commensal [29, 55]. Our additional analyses focusing on genes associated with specific functions show that protectins, and to a lesser extent adhesins and iron acquisition systems, are positively correlated with residence, while the signal was unclear for toxins. Interestingly, while the most resident phylogroups were identified as the most likely to generate extra-intestinal infections, better colonizing phylogroups such as B1, A.1 or E have been associated with pathovars causing severe and acute diarrhea [56]. These pathovars may benefit more from the ability to colonize rather than from the ability to reside.

Our results might also have implications for the evolution of *E. coli* in response to public health measures. On the one hand, 'multiplicative' measures increasing overall clearance rate (or reducing overall transmission) rate by a given factor are not predicted to alter the fitness hierarchy between phylogroups. On the other hand, 'additive' measures creating new sources of clearance, such as antibiotics, should favor the best colonizers at the expense of the residents. This could explain the resilience of the microbiota in settings with high antibiotic use [57]. Conversely, a reduction in disturbances of the microbiota could explain an increase in the frequency of phylogroup B2, as observed in France [25, 48]. The disproportionate impact of antibiotics on the most resident phylogroups could favor less pathogenic ones, which are better colonizers than residents. It could also generate stronger selection for antibiotic resistance among residents [58]. Yet, we found no link between residence and the presence of resistance genes in our data [25, 48]. This differs, for example, from *S. pneumoniae*, another human commensal bacteria and opportunistic pathogen, where serotypes carried for longer are also more antibioresistant [58].

Our study has limitations. We focused on colonization and clearance rates, which likely result from many underlying mechanisms, such as resource uptake, bacterial warfare or predation by phages [59]. Moreover, residence as measured in our study corresponds to the ability for a clone to maintain at densities high enough to be detected in the dominant flora. The dynamics of subdominant clones remains unobserved and would be interesting to quantify in further studies. Increasing the number of clones characterized by sample is expected to improve the completeness of the diversity observed in the data, but at a cost, for instance on the number of hosts included. Our results are supported by congruent analyses on two complementary datasets, focusing respectively on a large number of hosts (French) or on a large number of clones characterized per sample (USA). To test whether sampling a limited number of clones could bias our results or limit power, we randomly sub-sampled the USA dataset to five clones per sample (see Supporting information). Neither the trade-off nor the niche differentiation were biased by sub-sampling. The trade-off was still robustly identified, with a consistent negative correlation between residence and colonization similar to the one observed with the complete data, while the statistical power to detect niche differentiation was reduced, presumably because of the reduced diversity in sub-sampled data.

Colonization encompasses all events between the shedding of a clone by a host and its detection at high enough densities in another host. In our model, this series of events is encapsulated in the colonization rate. This is a valid description if the time spent by bacterial cells in the environment between hosts is short. In this model, better survival in the environment would result in a higher colonization rate. However, our model cannot account for long residence time in the environment or non-human hosts, which are known to harbor distinct *E. coli* clones [60, 61] with dynamics potentially largely independent from the human hosts. We investigated how transmission of clones from an external environmental reservoir would affect the outcome of our model (see Supporting information). Results show that an external reservoir could stabilize coexistence by providing a potential additional source of rare strains in the population.

In conclusion, by analyzing longitudinal data on *E. coli* epidemiological dynamics in healthy adults from two separate datasets, we discovered general principles structuring the diversity of this species: major lineages align along a colonization-residence trade-off equalizing their fitness, and their differentiated niches in humans stabilize their long-term coexistence. We also discovered a strong correlation between residence and pathogenicity. Our results on high-income countries, where the phylogroup with the slowest turnover B2 is over-represented [9, 43, 62], would be nicely complemented by longitudinal analyses in other countries with radically different phylogroup frequencies. Finally, our work also suggests mechanisms by which the micro-evolutionary dynamics of *E. coli* can be altered, and opens perspective for the study of epidemiological and evolutionary dynamics of other human commensal bacteria and opportunistic pathogens.

Material and Methods

Longitudinal datasets

We used two longitudinal datasets collected from adults living in high-income countries to perform the survival analyses. The first one, referred to as the French dataset, was original to this study. See the Supporting information for further details about the data collection and Fig S9 for a graphical representation of the longitudinal data. This dataset included 415 clones collected from 46 healthy subjects in the Paris area over the course of 10 samples approximately every two weeks (14.9 days on average between samples, SD = 4.9 days) in 2021 and 2022. The clones collected were differentiated using multilocus variable number tandem repeat analysis (MLVA) [63]. Each clone was assigned to one of twelve phylogroups, following Clermont et al. [21, 22]: A.1, A.2, B1, B2.1, B2.2, B2.3, C, D.1, D.2, E, F and G.

The second dataset, referred to as the USA dataset, was previously analyzed by Martinson et al. [44]. See Fig S8 for a graphical representation of the longitudinal data. This dataset included 117 clones collected from eight healthy subjects over periods ranging from 245 to 849 days (11.7 days on average between two samples, SD = 11.1 days). The clones collected were differentiated using GTG5 repetitive element amplification [64]. Each clone was assigned to one of nine phylogroups following the same method as for the French dataset: A.1, A.2, B1, B2.1, B2.2, B2.3, D (which includes D.1 and D.2 as defined in the French dataset), E and F. Phylogroups C and G did not appear in this dataset, respectively because a single clone of group C was recorded and was discarded, and because group G was not defined separately in the original study.

To assess whether our conclusions hold for colonization dynamics in infants, we considered a third dataset previously analyzed by Ostblom et al. [36] and referred to as the infant dataset. See the Supporting information for details about this dataset and the analyses performed, and Fig S10 for a graphical representation of the longitudinal data. This dataset included 273 clones collected from 130 infants in Sweden at fixed times since birth (after 3 days, 1, 2 and 4 weeks, 2, 6 and 12 months). The clones were assigned to one of the four main phylogroups: A (including A.1 and A.2), B1, B2 (including B2.1, B2.2, B2.3) and D (including D.1 and D.2).

Survival analyses

We used the software *R* [65] and the package *survival* [66] to perform survival analyses on the French and USA datasets. We used parametric accelerated failure time (AFT) models, which described the relationship between the time to an event – colonization or clearance – and a set of explanatory factors [67]. This type of model takes into account interval-censored data, i.e. the fact that colonization or clearance

can occur between two samples. We counted the residence time of a clone (time to clearance) starting from its first appearance in the host. We counted the time to colonization from the first appearance of the last preceding colonizing clone. For clarity, the results are presented in terms of the colonization rate, defined as the inverse of this time to colonization. We corrected the times to colonization to account for the differences in colonization rate due to differences in phylogroup frequency in the general population, based on the frequencies of the phylogroup in France and in the USA. See the Supporting information for more details on the method. We excluded left-censored data, but included right-censored data (clones still present in the last sample of the host) for the AFT models of residence time.

We considered how colonization and clearance depended on two qualitative factors – the host and the phylogroup of the focal clone – as well as the following quantitative factors: the total cell density of *E. coli* (in CFU/g), the number of clones inhabiting the host (other than the focal clone), and the number of other clones of the same phylogroup (other than the focal clone). We also considered the cell density of the clone considered for the AFT models of residence time of the USA dataset, as this information was not available for the French dataset owing to the smaller number of colonies typed per time-point. Because of the wide variations in cell densities observed between samples, we used their $\log_{10}$-transformed values.

We performed model selection based on Akaike's information criterion [AIC, 45] to identify the AFT model best explaining the residence times and times to colonization in the French and USA datasets. See the Supporting information for details about the model selection. In each case, we considered the models including any combinations of the explanatory factors presented above, with the time to the event following one of the three following distributions: exponential, Weibull and log-logistic. We ranked these models based on their ΔAIC, i.e. the difference between their own AIC and the smallest one, and kept all models with $\Delta\text{AIC} < 2$. We selected the most parsimonious model from this subset, unless a likelihood ratio test indicated that a more complex one from this subset explained the data significantly better. When discriminating between models with different distributions for the time to the event, we considered the exponential distribution to be the most parsimonious, as it assumed a constant hazard over time.

For both the French and USA datasets, we computed the estimate and standard error of median residence time and colonization rate of each phylogroup according to the AFT models. We defined the median colonization rate as the inverse of the estimated median time to colonization. To obtain estimates of typical colonization rates and residence times for each phylogroup, we computed these estimates for an average host based on those included in the dataset, and for the average values of all the quantitative explanatory factors included in the model. These estimates were used to compute the Spearman's correlation coefficient between colonization rate and residence time. We used the associated standard errors of parameter estimates to generate confidence intervals by parametric bootstrapping. We randomly sampled these distributions to generate one million sets, including each one residence time and one colonization rate per phylogroup. We computed the Spearman's correlation coefficient for each set, thus generating a distribution of the correlation between residence and colonization. We considered a correlation to be significant if 95% of the distribution of Spearman's coefficients was different from zero.

Correlation between residence and pathogenicity or the number of virulence-associated genes

We used an additional cross-sectional dataset, previously analyzed by Burgaya et al. [49], to estimate pathogenicity and the number of virulence-associated genes (VAG) of each phylogroup. It included 361 commensal and 746 pathogenic clones of *E. coli* (causing bloodstream infections), collected between 2000 and 2017 in Paris and the North of France. These clones were assigned to the same phylogroups as the ones described in the French dataset. These data also included the pathogenic status of each clone (1 for

pathogenic, 0 for commensal), as well as the number of VAG carried by the commensal clones, out of a list of 368 genes. Among those genes, some were specifically associated with one of the following type of virulence factors classically associated with pathogenic *E. coli* [46, 47, 68]: toxins (149 genes), adhesin (115 genes), iron acquisition systems (50 genes) and protectins (10 genes).

We estimated the pathogenicity of each phylogroup based on this additional dataset, using a binomial regression with the pathogenic status as a response variable and the phylogroup as the explanatory factor. We used the odds ratio for the association between infection and phylogroup, as estimated from this statistical model, as a proxy for pathogenicity. We also estimated the number of VAG (in total and of each type) using a binomial regression, with the proportion of VAG carried by the clone among those listed as the response variable and the phylogroup as the explanatory factor. We computed the estimated average number of VAG by multiplying the proportion predicted by the model by the total number of genes listed in the data.

As for residence and colonization (see above), we computed the estimates and standard errors for pathogenicity and the numbers of VAG per phylogroup. We used these distributions to generate confidence intervals by parametric bootstrapping. We sampled the distributions to generate one million set of each response variable (pathogenicity, total number of VAG and number of VAG associated with each type of factor) for each phylogroup. We computed a Spearman's correlation coefficients between each of those variables and the sets of residence times generated earlier (see above). We considered a correlation to be significant if 95% of the distribution of Spearman's coefficients was different from zero.

Description of the model of E. coli dynamics

We developed a continuous time, discrete-state Markov chain model describing the dynamics of *E. coli* clones in a host population. In this model, each host could contain between 0 and N clones, which belonged to one of G possible phylogroups. Hence, the discrete states of the model were the composition of the hosts in *E. coli* clones, expressed as unordered multisets of the phylogroups of the clones they contain. Hosts could be in one of $M = \binom{G+N}{G}$ states, which were numbered as follows:

$$\begin{aligned}
&S_1 = \emptyset \\
&S_2 = \{1\} \\
&S_3 = \{2\} \\
&\vdots \\
&S_{G+2} = \{1,1\} \\
&S_{G+3} = \{1,2\} \\
&\vdots \\
&S_M = \{\overbrace{G,\ldots,G}^{\times N}\}
\end{aligned} .$$

Hence, state S_1 corresponded to a host with no clone, S_2 to one with a clone of phylogroup 1, S_{G+2} to one with two clones of phylogroup 1, S_M to one with N clones of phylogroup G. The variables of the model were the set of probabilities $P_i(t)$ that a host was in state S_i at time t. The frequency of phylogroup k in the host population at time t was defined as:

$$F_k(t) = \sum_{i=1}^{M} V_{i,k} P_i(t) \ , \ \ V_{i,k} = \begin{cases} 0 & \text{if } \ i = 1 \\ m_{i,k}/l_i & \text{if } \ i > 1 \end{cases} \tag{4}$$

with $m_{i,k}$ the number of clones of phylogroup k in state i and l_i the total number of clones in state i. $V_{i,k}$ then corresponded to the proportion of clones of phylogroup k shed by a host in state i.

The transition between states was defined by the matrix $\mathbf{T_t}$ of size $M \times M$, whose elements, noted $\tau_{i,j}(t) > 0$ for the transition between state i and state j, were either a clearance rate, noted μ_{k,n,n_k}, or a colonization rate, noted $\lambda_{k,n}(t)$. Both were defined following the formula for an AFT model with an exponential distribution, with a parameter dependence mirroring our findings from the two datasets:

$$\begin{aligned} \mu_{k,n,n_k} &= 1/e^{a_k + b(n-1) + b_s(n_k - 1)} \\ \lambda_{k,n}(t) &= F_k(t)/e^{c_k + dn} \end{aligned} , \tag{1}$$

with a_k and c_k phylogroup-specific parameters, b and d parameters describing the impact of the number of clones in the host n, and b_s describing the additional impact of the number of clones of the same phylogroup n_k. Competition between clones regardless of phylogroup was expressed either as $b < 0$ if the clearance rate increased with n or as $d > 0$ if the colonization rate decreased with n. Additional competition between clones of the same phylogroup is expressed as $b_s < 0$, only affecting the clearance rate. The parameters determining these rates were chosen in accordance with the results of the survival analyses.

The transition matrix $\mathbf{T_t}$ was sparse, as the transition rate $\tau_{i,j}(t) > 0$ only if state j could be reached from state i with a single colonization or clearance event. The rows of $\mathbf{T_t}$ summed to 0, as the diagonal element $\tau_{i,i}(t) = -\sum_{j \neq i} \tau_{i,j}(t)$. We computed the matrix exponential $\mathbf{A_t} = e^{\mathbf{T_t}}$ using the package *expm* [69] to get the transition probabilities between states accounting for potential multiple transitions over a single time-step.

We simulated variations in state probabilities in discrete time with a time-step of a day, by updating the elements of the transition matrix $\mathbf{T_t}$ using the phylogroup frequencies of the previous day $F_k(t-1)$. In contrast to clearance rates, colonization rates were indeed not constant over the course of the simulation because they depended on the frequency of each phylogroup in the population. We stopped the simulation when a steady state was reached, i.e. for t^* such that $\sum_{i=1}^{N} |P_i(t^*) - P_i(t^* - 1)| < 10^{-12}$. We considered this steady state as the asymptotic state of our system, with the asymptotic probability of state S_i noted P_i^*.

We used this model to investigate the possible coexistence of several bacterial phylogroups in the population, and the impact of public health measures altering the colonization and clearance rates on the hierarchy of phylogroups. See Supporting information for additional analytical results developped to help interpret the simulation results.

Acknowledgements

This project has received funding from the CNRS (Momentum grant to F.B.), the French Foundation for Medical Research (Fondation pour la Recherche Médicale, to E.D. and O.Cl) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 949208 to F.B.). We thank Loubna Alavoine, Naima Beldjoudi, Charles Burdet, Corinne Da Costa Ribeiro Coutinho, Isabelle Gorenne, Milica Mandic, Ania Mlazindrou, and Valérie Vignali for their help setting up the French clinical study. We thank Cloé Sénéchal for her help with phylogroup typing, and Françoise Chau for her help processing the stool samples. We thank all volunteers for their participation in the studies.

Conflicts of interest

The authors declare no financial conflict of interest with the content of this article.

Author contributions

T.M-J and F.B. designed and conceptualized the project, with inputs from S.L.. R.A., O.Cl., S.D., C.F., J.M., P.R., X.D., F.L.N., S.W., E.D. and F.B. generated the data. T.M-J. analyzed the data with the help of O. Co. and F.B.. T.M-J. and F.B. developed the model, with the help of S.L.. T.M.-J. led the writing of the article, with inputs from all the co-authors.

References

[1] Christophe Fraser, William P Hanage, and Brian G Spratt. Recombination and the nature of bacterial speciation. *Science*, 315(5811):476–480, 2007. doi: 10.1126/science.1127573.

[2] Christophe Fraser, Eric J Alm, Martin F Polz, Brian G Spratt, and William P Hanage. The bacterial species challenge: Making sense of genetic and ecological diversity. *Science*, 323(5915):741–746, 2009. doi: 10.1126/science.1159388.

[3] Paul E Turner, Valeria Souza, and Richard E Lenski. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology*, 77(7):2119–2129, 1996. doi: 10.2307/2265706.

[4] Thomas Ferenci. Trade-off mechanisms shaping the diversity of bacteria. *Trends Microbiol*, 24(3): 209–223, 2016. doi: 10.1016/j.tim.2015.11.009.

[5] Fernando Baquero, Teresa M Coque, Juan Carlos Galán, and Jose L Martinez. The origin of niches and species in the bacterial world. *Front Microbiol*, 12:657986, 2021. doi: 10.3389/fmicb.2021.657986.

[6] Andrew J Sieben, Joseph R Mihaljevic, and Lauren G Shoemaker. Quantifying mechanisms of coexistence in disease ecology. *Ecology*, 103(12):e3819, 2022. doi: 10.1002/ecy.3819.

[7] Sarah Cobey and Marc Lipsitch. Niche and neutral effects of acquired immunity permit coexistence of pneumococcal serotypes. *Science*, 335(6074):1376–1380, 2012. doi: 10.1126/science.1215947.

[8] William P Hanage, Brian G Spratt, Katherine ME Turner, and Christophe Fraser. Modelling bacterial speciation. *Phil Trans R Soc B*, 361(1475):2039–2044, 2006. doi: 10.1098/rstb.2006.1926.

[9] Olivier Tenaillon, David Skurnik, Bertrand Picard, and Erick Denamur. The population genetics of commensal Escherichia coli. *Nat Rev Microbiol*, 8(3):207–217, 2010. doi: 10.1038/nrmicro2298.

[10] Karen L Kotloff, James P Nataro, William C Blackwelder, Dilruba Nasrin, Tamer H Farag, Sandra Panchalingam, Yukun Wu, Samba O Sow, Dipika Sur, and Robert F Breiman. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): A prospective, case-control study. *Lancet*, 382(9888):209–222, 2013. doi: 10.1016/S0140-6736(13)60844-2.

[11] Karen L Kotloff, James A Platts-Mills, Dilruba Nasrin, Anna Roose, William C Blackwelder, and Myron M Levine. Global burden of diarrheal diseases among children in developing countries: Incidence, etiology, and insights from new molecular diagnostic techniques. *Vaccine*, 35(49):6783–6789, 2017. doi: 10.1016/j.vaccine.2017.07.036.

[12] Lixin Zhang and Betsy Foxman. Molecular epidemiology of Escherichia coli mediated urinary tract infections. *Front Biosci*, 8(5):235–244, 2003. doi: 10.2741/1007.

[13] Betsy Foxman. The epidemiology of urinary tract infection. *Nat Rev Urol*, 7(12):653–660, 2010. doi: 10.1038/nrurol.2010.190.

[14] Marlieke EA De Kraker, Peter G Davey, Hajo Grundmann, and BURDEN Study Group. Mortality and hospital stay associated with resistant Staphylococcus aureus and Escherichia coli bacteremia: Estimating the burden of antibiotic resistance in Europe. *PLoS Med*, 8(10):e1001104, 2011. doi: 10.1371/journal.pmed.1001104.

[15] Michihiko Goto, Jennifer S McDanel, Makoto M Jones, Daniel J Livorsi, Michael E Ohl, Brice F Beck, Kelly K Richardson, Bruce Alexander, and Eli N Perencevich. Antimicrobial nonsusceptibility of gram-negative bloodstream isolates, Veterans Health Administration system, United States, 2003–2013. *Emerg Infect Dis*, 23(11):1815, 2017. doi: 10.3201/eid2311.161214.

[16] Li Liu, Hope L Johnson, Simon Cousens, Jamie Perin, Susana Scott, Joy E Lawn, Igor Rudan, Harry Campbell, Richard Cibulskis, and Mengying Li. Global, regional, and national causes of child mortality: An updated systematic analysis for 2010 with time trends since 2000. *Lancet*, 379(9832): 2151–2161, 2012. doi: 10.1016/S0140-6736(12)60560-1.

[17] JK Abernethy, AP Johnson, R Guy, N Hinton, EA Sheridan, and RJ Hope. Thirty day all-cause mortality in patients with Escherichia coli bacteraemia in England. *Clin Microbiol Infect*, 21(3): 251–e1, 2015. doi: 10.1016/j.cmi.2015.01.001.

[18] Sarah F Feldman, Elizabeth Temkin, Liat Wullfhart, Amir Nutman, Vered Schechner, Pnina Shitrit, Racheli Shvartz, Mitchell J Schwaber, Antoine Andremont, and Yehuda Carmeli. A nationwide population-based study of Escherichia coli bloodstream infections: Incidence, antimicrobial resistance and mortality. *Clin Microbiol Infect*, 28(6):879–e1, 2022. doi: 10.1016/j.cmi.2021.12.009.

[19] Marie Touchon, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, Stéphane Bonacorsi, Christiane Bouchier, and Odile Bouvet. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet*, 5(1): e1000344, 2009. doi: 10.1371/journal.pgen.1000344.

[20] Marie Touchon, Amandine Perrin, Jorge Andre Moura De Sousa, Belinda Vangchhia, Samantha Burn, Claire L O'Brien, Erick Denamur, David Gordon, and Eduardo PC Rocha. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLoS Genet*, 16(6): e1008866, 2020. doi: 10.1371/journal.pgen.1008866.

[21] Olivier Clermont, Stéphane Bonacorsi, and Edouard Bingen. Rapid and simple determination of the Escherichia coli phylogenetic group. *Appl Environ Microbiol*, 66(10):4555–4558, 2000. doi: 10.1111/1758-2229.12019.

[22] Olivier Clermont, Julia K Christenson, Erick Denamur, and David M Gordon. The Clermont Escherichia coli phylo-typing method revisited: Improvement of specificity and detection of new phylogroups. *Environ Microbiol Rep*, 5(1):58–65, 2013. doi: 10.1111/1758-2229.12019.

[23] Jan P Meier-Kolthoff, Richard L Hahnke, Jörn Petersen, Carmen Scheuner, Victoria Michael, Anne Fiebig, Christine Rohde, Manfred Rohde, Berthold Fartmann, and Lynne A Goodwin. Complete genome sequence of DSM 30083T, the type strain (U5/41T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genom Sci*, 9(1):1–19, 2014. doi: 10.1186/1944-3277-9-2.

[24] Camila Carlos, Mathias M Pires, Nancy C Stoppe, Elayse M Hachich, Maria IZ Sato, Tânia AT Gomes, Luiz A Amaral, and Laura MM Ottoboni. Escherichia coli phylogenetic group determination and its application in the identification of the major animal source of fecal contamination. *BMC Microbiol*, 10(1):1–10, 2010. doi: 10.1186/1471-2180-10-161.

[25] Méril Massot, Anne-Sophie Daubié, Olivier Clermont, Francoise Jaureguy, Camille Couffignal, Ghizlane Dahbi, Azucena Mora, Jorge Blanco, Catherine Branger, and France Mentré. Phylogenetic, virulence and antibiotic resistance characteristics of commensal strain populations of Escherichia coli from community subjects in the Paris area in 2010 and evolution over 30 years. *Microbiology*, 162 (4):642, 2016. doi: 10.1099/mic.0.000242.

[26] Christian-Daniel Köhler and Ulrich Dobrindt. What defines extraintestinal pathogenic Escherichia coli? *J Med Microbiol*, 301(8):642–647, 2011. doi: 10.1016/j.ijmm.2011.09.006.

[27] Erick Denamur, Olivier Clermont, Stéphane Bonacorsi, and David Gordon. The population genetics of pathogenic Escherichia coli. *Nat Rev Microbiol*, 19(1):37–54, 2021. doi: 10.1038/s41579-020-0416-x.

[28] Agnes E Wold, Dominique A Caugant, Gunilla Lidin-Janson, Peter de Man, and Catharina Svanborg. Resident colonic Escherichia coli strains frequently display uropathogenic characteristics. *J Infect Dis*, 165(1):46–52, 1992. doi: 10.1093/infdis/165.1.46.

[29] Tony Le Gall, Olivier Clermont, Stéphanie Gouriou, Bertrand Picard, Xavier Nassif, Erick Denamur, and Olivier Tenaillon. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group Escherichia coli strains. *Mol Biol Evol*, 24(11):2373–2384, 2007. doi: 10.1093/molbev/msm172.

[30] Kenneth L Vosti. A prospective, longitudinal study of the behavior of serologically classified isolates of Escherichia coli in women with recurrent urinary tract infections. *J Infect*, 55(1):8–18, 2007. doi: 10.1016/j.jinf.2007.01.006.

[31] Xin Fang, Jonathan M Monk, Sergey Nurk, Margarita Akseshina, Qiyun Zhu, Christopher Gemmell, Connor Gianetto-Hill, Nelly Leung, Richard Szubin, and Jon Sanders. Metagenomics-based, strain-level analysis of Escherichia coli from a time-series of microbiome samples from a Crohn's disease patient. *Front Microbiol*, 9:2559, 2018. doi: 10.3389/fmicb.2018.02559/full.

[32] Jessica C Seidman, Lashaunda B Johnson, Joshua Levens, Harran Mkocha, Beatriz Muñoz, Ellen K Silbergeld, Sheila K West, and Christian L Coles. Longitudinal comparison of antibiotic resistance in diarrheagenic and non-pathogenic Escherichia coli from young Tanzanian children. *Front Microbiol*, 7:1420, 2016. doi: 10.3389/fmicb.2016.01420.

[33] Hayley J Brodrick, Kathy E Raven, Teemu Kallonen, Dorota Jamrozy, Beth Blane, Nicholas M Brown, Veronique Martin, M Estée Török, Julian Parkhill, and Sharon J Peacock. Longitudinal genomic surveillance of multidrug-resistant Escherichia coli carriage in a long-term care facility in the United Kingdom. *Genome Med*, 9(1):1–11, 2017. doi: 10.1186/s13073-017-0457-6.

[34] I Adlerberth, F Jalil, B Carlsson, L Mellander, LÅ Hanson, P Larsson, K Khalil, and AE Wold. High turnover rate of Escherichia coli strains in the intestinal flora of infants in Pakistan. *Epidemiol Infect*, 121(3):587–598, 1998. doi: 10.1017/S0950268898001484.

[35] Forough Nowrouzian, Bill Hesselmar, Robert Saalman, Inga-Lisa Strannegård, Nils Åberg, Agnes E Wold, and Ingegerd Adlerberth. Escherichia coli in infants' intestinal microflora: Colonization rate, strain turnover, and virulence gene carriage. *Pediatr Res*, 54(1):8–14, 2003. doi: 10.1203/01.PDR.0000069843.20655.EE.

[36] Anna Ostblom, Ingegerd Adlerberth, Agnes E Wold, and Forough L Nowrouzian. Pathogenicity island markers, virulence determinants malX and usp, and the capacity of Escherichia coli to persist

in infants' commensal microbiotas. *Appl Environ Microbiol*, 77(7):2303–2308, 2011. doi: 10.1128/AEM.02405-10.

[37] Tommi Mäklin, Harry A Thorpe, Anna K Pöntinen, Rebecca A Gladstone, Yan Shao, Maiju Pesonen, Alan McNally, Pål J Johnsen, Ørjan Samuelsen, Trevor D Lawley, et al. Strong pathogen competition in neonatal gut colonisation. *Nature Communications*, 13(1):7417, 2022.

[38] H Je Sears, Inez Brownlee, and John K Uchiyama. Persistence of individual strains of Escherichia coli in the intestinal tract of man. *J Bacteriol*, 59(2):293–301, 1950. doi: 10.1128/jb.59.2.293-301.1950.

[39] HJ Sears and Inez Brownlee. Further observations on the persistence of individual strains of Escherichia coli in the intestinal tract of man. *J Bacteriol*, 63(1):47–57, 1952. doi: 10.1128/jb.63.1.47-57.1952.

[40] Peter FM Teunis, Eric G Evers, Paul D Hengeveld, Cindy M Dierikx, Cornelia CCH Wielders, and Engeline van Duijkeren. Time to acquire and lose carriership of ESBL/pAmpC producing E. coli in humans in the Netherlands. *PloS One*, 13(3):e0193834, 2018. doi: 10.1371/journal.pone.0193834.

[41] Diana Calderón, Paúl A Cárdenas, Belen Prado-Vivar, Jay P Graham, and Gabriel Trueba. A longitudinal study of dominant E. coli lineages and antimicrobial resistance in the gut of children living in an upper middle-income country. *J Glob Antimicrob Res*, 29:136–140, 2022. doi: 10.1016/j.jgar.2022.03.002.

[42] Bénédicte Condamine, Thibaut Morel-Journel, Florian Tesson, Guilhem Royer, Mélanie Magnan, Aude Bernheim, Erick Denamur, François Blanquart, and Olivier Clermont. Strain intrinsic properties and environmental constraints together shape Escherichia coli dynamics and diversity over a twenty-year human gut time series. *bioRxiv*, pages 2024–02, 2024. doi: 10.1101/2024.02.21.581337.

[43] Patricia Escobar-Páramo, Karine Grenet, Arnaud Le Menac'h, Luc Rode, Emmanuelle Salgado, Christine Amorin, Stéphanie Gouriou, Bertrand Picard, Mohamed Chérif Rahimy, and Antoine Andremont. Large-scale population structure of human commensal Escherichia coli isolates. *Appl Environ Microbiol*, 70(9):5698–5700, 2004. doi: 10.1128/AEM.70.9.5698-5700.2004.

[44] Jonathan NV Martinson, Nicholas V Pinkham, Garrett W Peters, Hanbyul Cho, Jeremy Heng, Mychiel Rauch, Susan C Broadaway, and Seth T Walk. Rethinking gut microbiome residency and the Enterobacteriaceae in healthy human adults. *ISME J*, 13(9):2306–2318, 2019. doi: 10.1038/s41396-019-0435-7.

[45] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans Automat Contr*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

[46] James R Johnson. Virulence factors in Escherichia coli urinary tract infection. *Clin Microbiol Rev*, 4(1):80–128, 1991. doi: 10.1128/cmr.4.1.80.

[47] James R Johnson and Thomas A Russo. Molecular epidemiology of extraintestinal pathogenic Escherichia coli. *EcoSal Plus*, 8(1):10–1128, 2018. doi: 10.1128/ecosalplus.esp-0004-2017.

[48] Julie Marin, Olivier Clermont, Guilhem Royer, Mélanie Mercier-Darty, Jean Winoc Decousser, Olivier Tenaillon, Erick Denamur, and François Blanquart. The population genomics of increased virulence and antibiotic resistance in human commensal Escherichia coli over 30 years in France. *Appl Environ Microbiol*, 88(15):e00664–22, 2022. doi: 10.1128/aem.00664-22.

[49] Judit Burgaya, Julie Marin, Guilhem Royer, Benedicte Condamine, Benoit Gachet, Olivier Clermont, Francoise Jaureguy, Charles Burdet, Agnes Lefort, and Victoire de Lastours. The bacterial genetic determinants of Escherichia coli capacity to cause bloodstream infections in humans. *PLoS Genet*, 19(8):e1010842, 2023. doi: 10.1371/journal.pgen.1010842.

[50] K Totsuka. Studien über Bacterium coli. *Z Hyg Infektionskr*, 45:115–124, 1903. ISSN 0340-1782.

[51] Peter Chesson. Mechanisms of maintenance of species diversity. *Annu Rev Ecol Evol Syst*, 31(1): 343–366, 2000. doi: 10.1146/annurev.ecolsys.31.1.343.

[52] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, and Andrey P Anokhin. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012. ISSN 0028-0836. doi: 10.1038/nature11053.

[53] Forough L Nowrouzian, Agnes E Wold, and Ingegerd Adlerberth. Escherichia coli strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *J Infect Dis*, 191(7):1078–1083, 2005. doi: 10.1086/427996.

[54] FL Nowrouzian, AE Östblom, AE Wold, and I Adlerberth. Phylogenetic group B2 Escherichia coli strains from the bowel microbiota of Pakistani infants carry few virulence genes and lack the capacity for long-term persistence. *Clin Microbiol Infect*, 15(5):466–472, 2009. doi: 10.1111/j.1469-0691.2009.02706.x.

[55] James R Johnson, Connie Clabots, Stephen B Porter, Tricia Bender, Brian D Johnston, and Paul Thuras. Intestinal Persistence of Colonizing Escherichia coli Strains, Especially ST131-H 30, in Relation to Bacterial and Host Factors. *J Infect Dis*, 225(12):2197–2207, 2022. ISSN 0022-1899. doi: 10.1093/infdis/jiab638.

[56] Patricia Escobar-Páramo, Olivier Clermont, Anne-Béatrice Blanc-Potard, Hung Bui, Chantal Le Bouguénec, and Erick Denamur. A specific genetic background is required for acquisition and expression of virulence factors in Escherichia coli. *Mol Biol Evol*, 21(6):1085–1094, 2004. ISSN 1537-1719. doi: 10.1093/molbev/msh118.

[57] Vu Thi Ngoc Bich, Ngoc Giang Le, David Barnett, Jiyang Chan, Niels van Best, Tran Dac Tien, Nguyen Thi Hien Anh, Tran Huy Hoang, H Rogier van Doorn, and Heiman FL Wertheim. Moderate and transient impact of antibiotic use on the gut microbiota in a rural Vietnamese cohort. *Sci Rep*, 12(1):20189, 2022. doi: 10.1038/s41598-022-24488-9.

[58] Sonja Lehtinen, François Blanquart, Nicholas J Croucher, Paul Turner, Marc Lipsitch, and Christophe Fraser. Evolution of antibiotic resistance is linked to any genetic mechanism affecting bacterial duration of carriage. *Proc Natl Acad Sci*, 114(5):1075–1080, 2017. doi: 10.1073/pnas.1617849114.

[59] Jonathan NV Martinson and Seth T Walk. Escherichia coli residency in the gut of healthy human adults. *EcoSal Plus*, 9(1), 2020. doi: 10.1128/ecosalplus.ESP-0003-2020.

[60] P Money, AF Kelly, SWJ Gould, J Denholm-Price, EJ Threlfall, and MD Fielder. Cattle, weather and water: Mapping Escherichia coli O157: H7 infections in humans in England and Scotland. *Environ Microbiol*, 12(10):2633–2644, 2010. doi: 10.1111/j.1462-2920.2010.02293.x.

[61] Catherine Ludden, Kathy E Raven, Dorota Jamrozy, Theodore Gouliouris, Beth Blane, Francesc Coll, Marcus de Goffau, Plamena Naydenova, Carolyne Horner, and Juan Hernandez-Garcia. One health genomic surveillance of Escherichia coli demonstrates distinct lineages and mobile genetic elements in isolates from humans versus livestock. *MBio*, 10(1):e02693–18, 2019. doi: 10.1128/mbio.02693-18.

[62] Nancy de Castro Stoppe, Juliana S Silva, Camila Carlos, Maria IZ Sato, Antonio M Saraiva, Laura MM Ottoboni, and Tatiana T Torres. Worldwide phylogenetic group patterns of escherichia coli from commensal human and wastewater treatment plant isolates. *Frontiers in microbiology*, 8: 2512, 2017.

[63] François Caméléna, André Birgy, Yasmine Smail, Céline Courroux, Patricia Mariani-Kurkdjian, Simon Le Hello, Stéphane Bonacorsi, and Philippe Bidet. Rapid and simple universal Escherichia coli genotyping method based on multiple-locus variable-number tandem-repeat analysis using single-tube multiplex PCR and standard gel electrophoresis. *Appl Environ Microbiol*, 85(6):e02812–18, 2019. doi: 10.1128/AEM.02812-18.

[64] BR Mohapatra and A Mazumder. Comparative efficacy of five different rep-PCR methods to discriminate Escherichia coli populations in aquatic environments. *Water Sci Technol*, 58(3):537–547, 2008. doi: 10.2166/wst.2008.424.

[65] RCoreTeam. R: A language and environment for statistical computing. Vienna, Austria, 2018.

[66] Terry M Therneau. A Package for Survival Analysis in R. Vienna, Austria, 2022.

[67] Lee-Jen Wei. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Stat Med*, 11(14-15):1871–1879, 1992. doi: 10.1002/sim.4780111409.

[68] Marco Galardini, Olivier Clermont, Alexandra Baron, Bede Busby, Sara Dion, Sören Schubert, Pedro Beltrao, and Erick Denamur. Major role of iron uptake systems in the intrinsic extra-intestinal virulence of the genus Escherichia revealed by a genome-wide association study. *PLoS Genet*, 16 (10):e1009065, 2020. doi: 10.1371/journal.pgen.1009065.

[69] Vincent Goulet, Christophe Dutang, Martin Maechler, David Firth, Marina Shapira, and Michael Stadelmann. Expm: Matrix Exponential, Log, 'etc'. Vienna, Austria, 2021.