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], Olivier Clermont[d], Sara Dion[d], Clarisse Figueroa[d], Jonathan Martinson[e], Pascal Ralaimazava[f], 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 commensal bacteria and major pathogen *Escherichia coli*, much less is known about its ecological and evolutionary dynamics in populations of 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 trade-off between the ability to establish in a new host (colonization) and the ability to remain in the host (residence). Hence, 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 within the host. 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 such as strains of the extra-intestinal pathogenic phylogroup B2.3.

Introduction

Escherichia coli is a commensal bacteria of the human gut [1] and an opportunistic pathogen, responsible for intestinal [2, 3], urinary tract [4, 5] and bloodstream infections [6, 7], causing around one million deaths worldwide each year [8, 9, 10]. There is no strict phylogenetic separation between commensal and extraintestinal pathogenic *E. coli* [11], the transition to infection being the result of complex and hardly predictable interactions between several genetic events [12]. Genes associated with pathogenicity (the probability to cause an infection) and virulence (the severity of infection) could evolve as a result of their role in commensalism [13, 14]. Studying the dynamics of commensal *E. coli* is therefore essential to better understand the evolution of pathogenicity and virulence.

As a species, *E. coli* is not only highly diverse, but also highly structured [15, 16], notably around lineages called 'phylogroups' [17, 18], which are largely monophyletic [1, 19] and differ in their virulence and antibiotic resistance profiles [20, 21]. At a finer scale, strains are structured by sequence types [16], which also differ in their characteristics and are congruent with phylogroups [22]. Understanding how this structure has arisen and been maintained is essential for predicting the future evolution of commensal strains of *E. coli*. This clear phylogenetic structure echoes the long-standing question of bacterial diversification under the combined action of mutation and recombination [23, 24]. Multiple mechanisms have been proposed to explain the observed diversity and coexistence of different bacterial strains. A recombination rate negatively correlated with genetic distance can structure diversity, leading to genetically distinct groups even without selection [23, 25]. The diversity of ecological niches favors genetic diversity [26, 27]. For example, the combination of a specific immune response and a general host immune response promotes coexistence between serotypes of *Streptococcus pneumoniae* [28]. Trade-offs between life-history traits can also enable species to diversify by adopting distinct strategies [29, 30].

Characterizing such life-history traits among different *E. coli* strains and their impact on coexistence requires characterizing their epidemiological dynamics with longitudinal sampling of the same hosts over time. Such studies have been performed with diseased individuals [e.g. 31, 32], notably to assess the impact of antibiotic use [33, 34], but remain rare in healthy individuals, in part due to the complexity of their implementation. Among studies of healthy individuals, some focus on the special case of initial colonization of infants during their first months of life [35, 36, 37, 38]. Other studies on adults include a limited number of data points [39, 40, 41, 42].

In this study, we investigated the colonization dynamics of *E. coli* strains belonging to different phylogroups using two remarkably detailed longitudinal sampling of the gut microbiome of healthy adult subjects: an original dataset collected in France, and a dataset previously collected in the USA and analyzed by Martinson et al. [43]. We find that a trade-off between colonization and residence, together with niche differentiation between phylogroups, can explain the coexistence of multiple strains among commensal bacteria.

Results

Survival analyses

We studied the turnover dynamics of commensal *E. coli* using two datasets with longitudinal sampling over several months and accurate identification of clones. The first one (called thereafter the 'French dataset') includes 415 clones sampled from 46 hosts over periods of 111 to 135 days. The second one (called thereafter the 'USA dataset') includes 117 clones sampled from 8 hosts over periods of 245 to 849 days. In both studies, clones were assigned to phylogroups with the Clermont method and differentiated within a single individual by additional PCR-based DNA fingerprinting techniques (see Material and

Methods). [17, 18].

Survival models, specifically parametric accelerated failure time (AFT) model, were used to assess the impact of various factors on the rate at which clones colonized new hosts or were cleared. The response variable used in the models was either the colonization rate of the clones, i.e. the rate at which a new clone was detected for the first time in a host, or their residence time in the gut, i.e. the number of days between the first and last detection of a clone in a host. The potential explanatory factors considered were the host, phylogroup, the total number of clones and the number of clones of the same phylogroup in the host, and the total cell density. The cell density of the clone itself was also considered for the AFT model of the residence time in the USA data only, as no precise estimate was available for the French data. Model selection based on Akaike's information criterion [AIC, 44] was performed to identify the explanatory factors included in the best model (Supporting information).

Determinants of the residence time

The survival analyses performed on both datasets highlighted differences in residence times between hosts and phylogroups (see Table S2 for the French dataset and Table S4 for the USA dataset). Both datasets also showed evidence for niche differentiation between phylogroups: the residence of a clone in a host was more strongly reduced by the presence of clones if they belonged to the same phylogroup. The analysis of the French dataset showed that residence time was reduced by the presence of other clones in the host ($p < 0.0001$), even more if those clone were of the same phylogroup as the one considered ($p = 0.0079$). According to the model, 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. The analysis of the USA dataset also showed reduced residence times when other clones were in the host ($p = 0.0006$), with an additional effect of those of the same phylogroup ($p = 0.0328$). 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. As expected, longer residence times were associated with higher density of the clone considered ($p = 0.0473$), with a residence time multiplied by 1.32 (95%CI $= [1.002, 1.499]$) when the clone density was multiplied by 10. Additional analyses using semi-parametric Cox models with the same factors showed similar relationships, confirming the robustness of these results to the type of survival model used (Supporting information)

Determinants of the colonization rate

Whereas residence was expected to depend solely on dynamics within the host, the colonization rate was also expected to be influenced by the relative frequencies of phylogroups in the host population in a simple model of inter-human transmission. To account for this, we corrected the colonization rates observed in the two datasets by the relative frequencies of phylogroups in their respective countries, based on cross-sectional external data (Supporting information). Thereafter we analyzed these corrected colonization rates.

As for residence times, the survival analyzes showed differences between hosts and phylogroups in their colonization rates (see Table S6 for the French dataset and Table S8 for the USA dataset). In the French dataset, colonization was more frequent in hosts already harboring a greater number of clones ($p < 0.0001$), with a predicted rate multiplied by 1.76 (95% $CI = [1.562, 1.970]$) for each additional clone. This was also the case in the USA dataset ($p < 0.0001$), with a predicted rate multiplied by 2.89 (95% $CI = [2.342, 3.589]$) for each additional clone. However, in neither dataset was the colonization rate influenced by the bacterial density in the host, nor the fact that the colonizer and residents belonged to the same phylogroup (factors not included in the best model). Additional analyses using Cox models

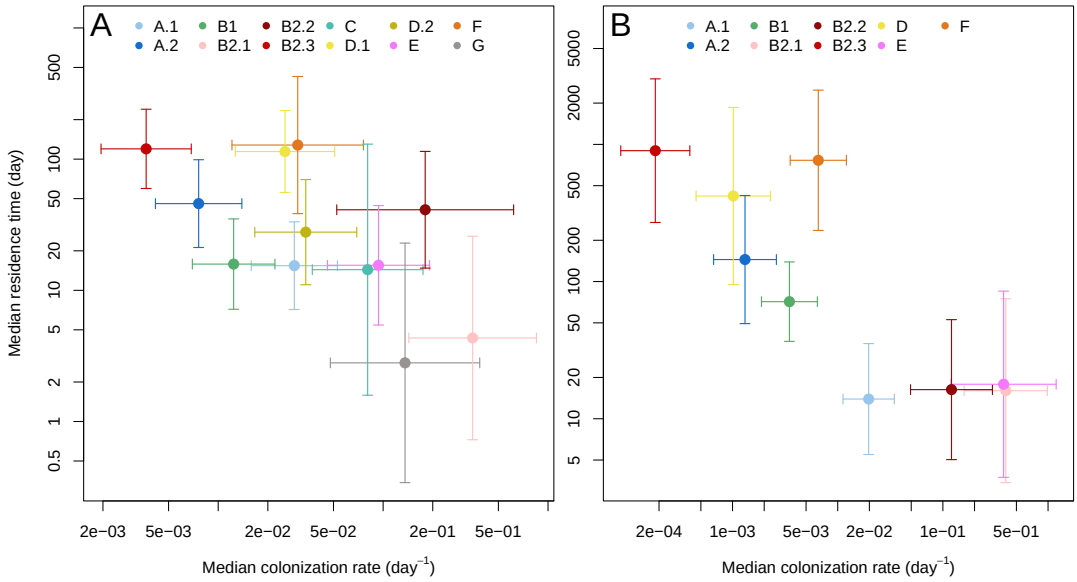

Figure 1: Predicted median residence time as a function of the predicted median colonization rate for each phylogroup appearing in the French (A) and the USA (B) dataset, for the average number of clones and densities observed in the data and the average host among those harboring a clone of the phylogroup at least once according to the data. The errors represent the 95% confidence interval of the predicted value.

with the same factors show similar relationships, again confirming the robustness of these results to the type of survival model used (Supporting information).

Trade-off between colonization and residence

Following analyses of both datasets, we compared the median colonization rate and residence time estimated for each phylogroup by the best AFT models. Phylogroups were distributed along an axis, ranging from strong colonizers with short residence times, to poor colonizers with long residence times (Fig. 1). To formally test the significance of the correlation between the two traits, we randomly drew one million sets of one colonization rate and one residence time for each phylogroup, in the distribution of their respective predicted values according to the models. The negative correlation was significant in both the French (Spearman's ρ 95%CI $= [-0.762, -0.238]$, $p = 0.0016$) and the USA (Spearman's ρ 95%CI $= [-0.923 - 0.599]$, $p < 0.0001$) datasets. This suggests the existence of a robust trade-off between the colonization and residence abilities of the phylogroups.

To further assess the robustness of the trade-off, we performed the same analysis with colonization rates corrected by the phylogroup frequencies provided by the two datasets themselves instead of external data (Supporting information). We again found a negative trade-off, both for the French (95%CI $= [-0.618, -0.832]$, $p = 0.0002$) and the USA dataset (95%CI $= [-0.900, -0.583]$, $p < 0.0001$). Besides, we also investigated this trade-off in additional, large longitudinal dataset of infants followed for one year after their birth in Sweden [37] (Supporting information). While this dataset focuses on the initial colonization of the gut instead of the adult microbiota, the trade-off between residence time and colonization rate also clearly appeared.

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) [45, 46]. Pathogenicity (the odds ratio for the association between bloodstream infections and phylogroup) and the number of VAG carried were estimated for each phylogroup, based on a dataset previously analyzed by Marin et al. [47] and Burgaya et al. [48]. The estimated pathogenicity of phylogroups was positively correlated with residence for both datasets (French dataset: 95%CI $= [0.105, 0.741]$, $p = 0.0131$); USA dataset: 95%CI $= [0.217, 0.833]$, $p = 0.0035$)(Fig 2A, 2B). The same data was used to estimate the average number of VAG in each phylogroup (Fig 2C). Their total number was positively correlated with residence for the French data (95%CI $= [0.049, 0.531]$, $p = 0.0221$) and the USA data (95%CI $= [-0.017, 0.533]$, $p = 0.0530$). Differences were observed in the strength of the correlation when considering genes associated with specific functions (Fig 2C): those linked to protectins were the most strongly correlated with residence, while those linked with toxins were not. Genes linked to iron acquisition ($p = 0.0473$) and to adhesins ($p = 0.0178$) were significantly correlated with residence, respectively for the French and the USA datasets.

Model of bacterial epidemiological dynamics

Following data analysis, we derived implications of our results concerning colonization and clearance rates for the epidemiological dynamics of *E. coli*. We developed a continuous time, discrete-state Markov chain model describing the dynamics of clones in a host population. In this model, discrete states are sets of the phylogroups of *E. coli* clones potentially present in the hosts. Phylogroups are numbered from 1 to G and states are numbered in increasing order of the number of clones in the host, so that an uncolonized host is in state $S_1 = \{\emptyset\}$, a host colonized by a single clone of phylogroup 2 in state, $S_2 = \{1\}$, $S_3 = \{2\}$, and a host colonized by two clones of group 1 in state $S_{G+1} = \{1, 1\}$. The variable of the model is the probability $P_i(t)$ that a host is in state S_i at time t, which can also be interpreted as the proportion of hosts in state S_i.

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

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

with a_k and c_k phylogroup-specific parameters, $F_k(t)$ the frequency of phylogroup k in the host population, n the number of distinct clones in the host and n_k the number of clones of phylogroup k in the host. The impacts of n on clearance and colonization rates are modulated by b and d respectively, with competition between hosts on residence or colonization corresponding to $b < 0$ and $d < 0$ respectively. Additional competition between clones of the same phylogroup is described by $b_s < 0$. These parameters correspond to those included in the best retained AFT statistical models. As niche differentiation was only evidenced for residence, no additional competition parameter was considered for colonization.

Phylogroup k is characterized by a colonization-to-clearance ratio $R_k = e^{a_k - c_k}$, which can be interpreted as its basic reproduction number, barring potential additional effects of immunity. Let us consider a single clone of phylogroup k introduced in a population of H hosts without any other clone, large enough such that the number of non-colonized hosts $H - 1 \approx H$. The frequency of this clone is therefore $F_k(t) = 1/H$ and its expected residence time in the colonized host is $1/\mu_{k,1,1}$. The expected number of hosts colonized from this single introduction is then:

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

Since a_k and c_k are the only parameters specific to phylogroup k, R_k also predicts the competitive hierarchy between phylogroups. Let us consider two phylogroups x and y, such that $R_x > R_y$. Without

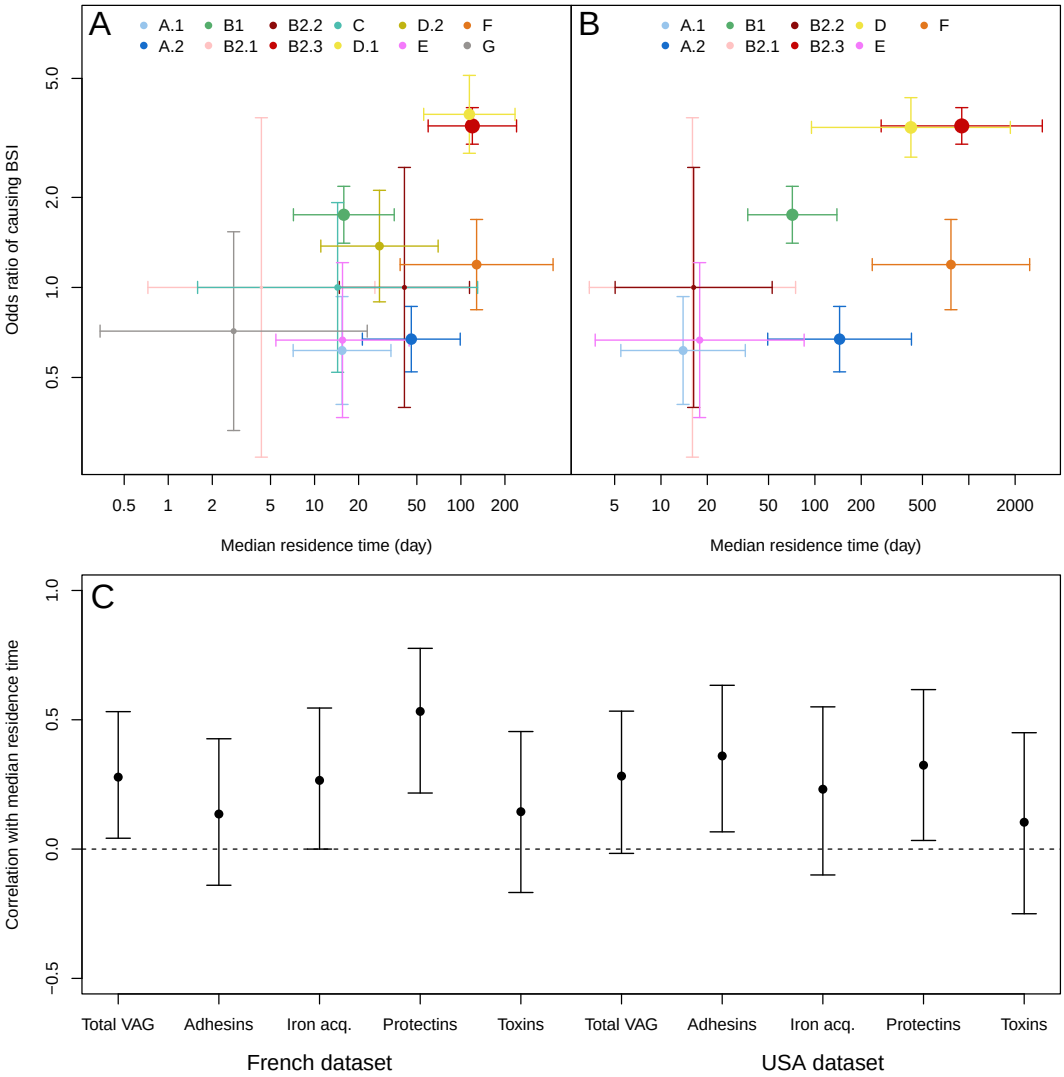

Figure 2: Correlations between residence and pathogenicity or virulence-associated genes (VAG). A and B: odds ratio for the association between bloodstream infections (BSI) and phylogroup with their 95% confidence interval (vertical), as a function of their median residence time predicted by our AFT models for the French (A) or USA dataset (B) with their 95% confidence interval (horizontal). The sizes of the dots correspond to the number of samples considered to estimate pathogenicity. C: 95% confidence intervals for the correlation between the number of VAG (either total or the subset of those associated with adhesins, iron acquisition systems, protectins or toxins) and the median residence time predicted by our AFT models (for the French or the USA dataset).

niche differentiation ($b_s = 0$), phylogroup x can invade a host population in which phylogroup y has reached an asymptotic frequency, but not the opposite (Supporting information), so that the phylogroup with highest R_k eventually prevails.

Impact of changes in colonization or clearance rates on bacterial evolution

Our model suggests that non phylogroup-specific public hygiene measures may impact the competitive hierarchy between phylogroups, depending on the way they change colonization or clearance rates. Let us consider two phylogroups x and y with $R_x > R_y$. On the one hand, multiplicative changes, i.e. variations of colonization or clearance rate by a constant factor, do not affect the hierarchy. For instance, a multiplicative change in clearance rate by a factor η, such that $\mu'_{k,n,n_k} = \eta\mu_{k,n,n_k}$, leads to a decrease in the colonization-to-clearance ratios by the same factor η, i.e. $R'_x = R_x/\eta$ and $R'_y = R_y/\eta$, so that $R'_x > R'_y$.

On the other hand, additive changes, i.e. inclusion or removal of other sources of colonization or clearance (e.g. antibiotics), can affect the hierarchy between phylogroups. For instance, including an additional clearance rate θ, such that $\mu''_{k,n,n_k} = \theta + \mu_{k,n,n_k}$, amounts to multiplying the clearance rate by a phylogroup-specific factor equal to:

$$\mu''_{k,n,n_k} = \theta^m_k \mu_{k,n,n_k} \ , \ \theta^m_k = \frac{\mu_{k,n,n_k} + \theta}{\mu_{k,n,n_k}}. \tag{3}$$

so that $R''_x = R_x/\theta^m_x$ and $R''_y = R_y/\theta^m_y$. This factor θ^m_k is negatively correlated with μ_{k,n,n_k}, therefore the longest-residing phylogroup are expected to be the most affected by this additional increase in clearance. Following Eq. 3, the additional clearance rate needed to cancel out the difference in ratio between phylogroups, i.e. such that $R''_x = R''_y$, is defined as:

$$\theta^{eq} = \frac{R_x - R_y}{R_y e^{a_x} - R_x e^{a_y}}. \tag{4}$$

It should be noted that $\theta^{eq} > 0$, only if $c_x > c_y$, i.e. an additional source of clearance can only negate the advantage of phylogroup x if its strategy is more residence-oriented than that of y, in which case the hierarchy is reverse if $\theta > \theta^{eq}$. These changes can also generate variations in frequencies of the clones, with clone y overtaking x when θ is high enough (Fig. 3).

Conditions for the coexistence of multiple phylogroups

We computed two indices quantifying the coexistence between phylogroups: the within-host diversity $E^W(t)$ and the inter-host diversity $E^I(t)$. $E^W(t) \in [0, 1]$ is the total proportion of hosts carrying clones from at least two different phylogroups. $E^I(t) \in [0, \ln(G)]$ is a Shannon index on the relative frequencies of the phylogroups in the host population:

$$E^I(t) = -\sum_{k=1}^{G} f^h_k(t) \ln\left(f^h_k(t)\right), \ \ f^h_k(t) = \frac{F^h_k(t)}{\sum_{k=1}^{G} F^h_k(t)}. \tag{5}$$

We considered a simple case of a couple of 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 $D_{i,j} = R_i - R_j$. Using discrete-time simulations of the model with a time step of a day, we computed the asymptotic within-host and inter-host coexistence values reached by this system. We considered that the simulations had reached their asymptotic state after t^* time-steps, such that $\sum_{i=1}^{N} |P_i(t^*) - P_i(t^* - 1)| < 10^{-12}$, i.e. when state probabilities essentially no longer varied from one time step to the next.

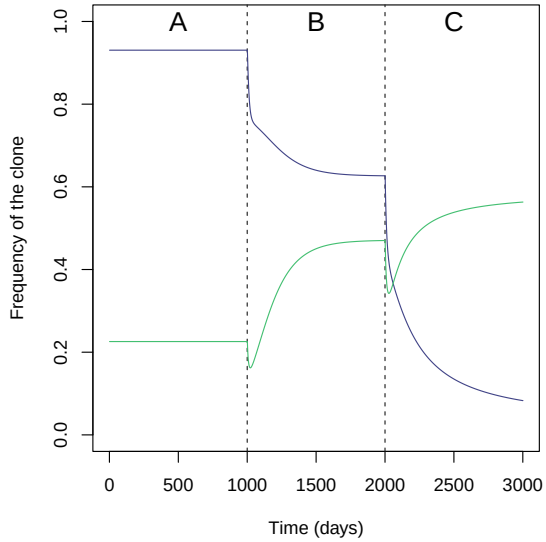

Figure 3: Example of variation in the frequencies of phylogroups x (purple) and y (green) over 3000 simulated days of the model, with $R_x = 3.5$, $R_y = 3$, $\theta_{eq} = 0.15$, $b = 0$, $b_s = -0.1$ and $d = 0$, . Three time-periods of 1000 days each (separated by dashed lines) are defined: (A) no additive clearance ($\theta = 0$) for $t \in [0, 1000[$, (B) $\theta = 0.1$ for $t \in [1000, 2000[$ and (C) $\theta = 0.2$ for $t \in [2000, 3000[$. Because clone x is less affected by the additive clearance, it becomes more frequent than clone y when $\theta > \theta_{eq}$ (C).

Results show that the trade-off alone allows only structurally unstable coexistence. (Fig 4). In line with the mathematical analysis of the model (Supporting information), coexistence is only maintained for the strains with exactly equal colonization-to-clearance ratios ($D_{i,j} = 0$) if there is no additional competition between clones of the same phylogroup ($b_s = 0$, see Eq. 1). However, coexistence can be stable if higher competition between clones of the same phylogroup creates additional clearance ($b_s < 0$). As b_s becomes more negative (greater niche differentiation), the range of values of $D_{i,j}$ for which coexistence is stable widens, but within-host diversity E^{W*} also declines (Fig 4A). This is because lower values of b_s lead to higher average clearance rates overall, thus decreasing the average number of clones per host and within-host diversity. This is not the case for inter-host diversity, whose maximal value remains the same when b_s varies (Fig 4B). The simulation results presented in Fig 4 correspond to $\overline{R}_{i,j} = 2$, $b = 0$ and $d = 0$, but similar results are obtained for other parameter values. As higher values of $\overline{R}_{i,j}$ increase the average basic reproduction rate of phylogroups, they also widen the set of $D_{i,j}$ for which coexistence is stable. The same applies to high values of b (clearance reduced as the number of clones increases) or low values of d (increased colonization with the number of clones), since these two parameters affect every phylogroup considered equally. By respectively reducing the overall clearance rates and increasing the overall colonization rates when multiple clones are present in the host, both increase the average number of clones per host, thus increasing diversity in the same way as $\overline{R}_{i,j}$.

Discussion

By analyzing longitudinal datasets on *E. coli* gut colonization dynamics, we identified two mechanisms explaining the stable coexistence of *E. coli* phylogroups. The first is the trade-off between the colonization and residence abilities across phylogroups. The 'resident' or 'transient' nature of *E. coli* clones, which has been observed for over 100 years [49, 40], is a property partly explained by phylogroups and inversely correlated with the ability to colonize hosts. This trade-off is equalizing *sensu* Chesson [50], meaning that it reduces differences in average fitness between phylogroups. In our model of colonization dynamics with human-to-human transmission, fitness is mathematically equal to the colonization-to-clearance ratio

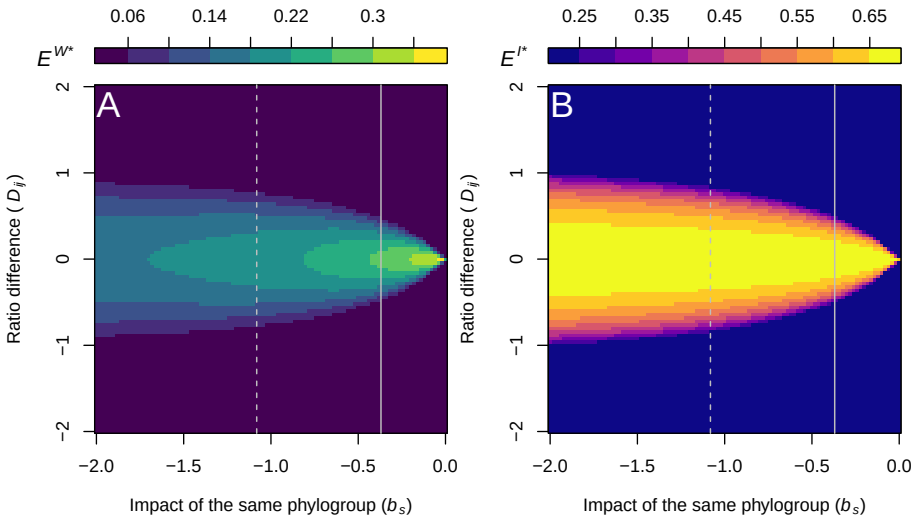

Figure 4: Asymptotic within host E^{W*} (A) and inter-hosts E^{I*} (B) diversity as a function of the ratio difference $D_{i,j}$ and the strength of clones of the same phylogroup on clearance b_s. Vertical lines correspond to the values of b_s estimated for the French (solid) and USA datasets (dashed).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

R_k. Therefore, the coexistence of multiple phylogroups in a population at equilibrium implies that these phylogroups compensate shorter residence times with a higher colonization rates (or vice versa) to reach achieve similar fitness. Indeed, clones with lower fitness could not rise in frequency, while clones with larger fitness would invade and replace all other clones.

The second mechanism is niche differentiation – a stabilizing mechanism explaining long-term coexistence [50]. In both datasets, residence was shorter when a clone inhabited a host together with another clone of the same phylogroup. This suggests stronger competition within phylogroups than between them that could be explained by a greater overlap in ecological niches. Interestingly, colonization did not depend on the number of clones of the same phylogroup, suggesting that the niche overlap only affects clones *after* the colonization of a new host.

Our analyses also highlighted substantial heterogeneity between hosts in bacteria dynamics, consistent with the absence of 'core resident microbiome' initially reported by Martinson et al. using the USA data [43]. 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 host's suitability for colonization or residence. They could also be triggered by the colonization of several clones in a short time-span, temporarily destabilizing the microbial community. These two explanations are not exclusive and could both contribute to temporal variability.

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 [51, 14, 52]. 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 to its residence ability as a commensal [14, 53]. Further analyses focusing on genes associated with specific functions suggests that protectins, and to a lesser extent adhesins and iron acquisition systems are positively correlated with residence, while this

was not the case 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 [54], which may benefit more from colonization abilities than residence ones.

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 competitive hierarchy between strains. On the other hand, 'additive' measures creating new sources of clearance, such as antibiotics, should favor the best colonizers at the expense of the resident clones. This could explain the resilience of the microbiota in settings with high antibiotic use [55]. Conversely, a reduction in disturbances of the microbiota could explain an increase in the frequency of phylogroup B2, as observed in France [21, 47]. The disproportionate impact of antibiotics on the most resident phylogroups could favor less pathogenic strains, which are better colonizers than residents. It could also generate stronger selection for antibiotic resistance among residents [56]. Yet, we found no link between residence and the presence of resistance genes [21, 47]. This differs, for example, from *S. pneumoniae*, another human commensal bacteria and opportunistic pathogen, where serotypes carried for longer are also more antibioresistant [56].

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 [57]. Moreover, residence as measured in our study actually 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. Similarly, colonization encompasses all events between the shedding of a clone by a host and its detection at high enough densities in another host. If the time spent by bacterial cells in the environment between hosts is short, our model will encapsulate this series of events in the colonization rate. Thus, a higher colonization rate could result from better survival in the environment. Our model cannot account for long residence time in the environment or non-human hosts, which are known to harbor distinct *E. coli* clones [58, 59], with potentially independent dynamics from that in the human population. We investigated how transmission of clones from an external environmental reservoir would affect the outcome of our model (Supporting information), and found that an external reservoir would stabilize coexistence, by providing an 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 datasets originating from two high-income countries, we discovered general principles structuring the diversity of this species: major clones aligned along a colonization-residence trade-off equalizing their fitness, and occupy distinct niches in humans, thus stabilizing their long-term coexistence. We also discovered a strong correlation between residence and pathogenicity. This work 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

Data

For the survival analyses, we use two datasets: an original dataset collected in Paris in 2022-2023 and en existing dataset collected in the USA and analyzed in Martinson et al. [43]. The data from France includes 415 clones collected from 46 healthy subjects in the Paris area over the course of 10 samplings approximately every two weeks (14.9 days on average between samples, SD = 4.9 days) in 2021 and

2022. Clones were assigned to one of twelve phylogroups following Clermont et al.[17, 18]: A.1, A.2, B1, B2.1, B2.2, B2.3, C, D.1, D.2, E, F and G. Different clones of the same phylogroups were differentiated using multilocus variable number tandem repeat analysis (MLVA) [60]. Further details on these data are given in Supplementary information. The data from Martinson et al. includes 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). Clones were assigned to nine phylogroups following the same method, with the following differences: D.1 and D.2 were considered altogether as D and groups C and G did not appear in the dataset, respectively because a single clone of group C was recorded, and because group G was not defined separately in the original study. Different clones of the same phylogroup in the same host were differentiated using GTG5 repetitive element amplification [61].

To assess pathogenicity and the number of virulence-associated genes (VAG), an additional cross-sectional dataset, previously analyzed by Burgaya et al. [48], was used. The data were collected between between 2000 and 2017 in Paris and the North of France, and includes 361 commensal and 746 pathogenic (causing bloodstream infections) clones of *E. coli*, which were assigned to the same phylogroups as those in the French dataset described above. The data includes the pathogenic status of each clone (1 for pathogenic, 0 commensal), as well as the number of VAG carried by the commensal clones, out of a list of 368. Some of these genes were more specifically associated with one of the following types of virulence factors classically associated with pathogenic *E. coli* [45, 46, 62]: toxins (149 genes), adhesin (115 genes), iron acquisition systems (50 genes) and protectins (10 genes).

Survival analyses

We used the software *R* [63] and the package *survival* [64] to perform survival analyses on both datasets. We used parametric accelerated failure time (AFT) models, which described the main relationship between time to an event – either colonization or clearance – and explanatory factors [65]. These models accommodated the interval-censored longitudinal data considered. We defined the time to colonization by a clone as the number of days between the last previous colonization and the first appearance of the this clone in the host, and its 'colonization rate' as the inverse of the time to colonization. We defined, the time to clearance of a clone, or 'residence time', as the number of days since its first appearance in the host. For the survival analyses, we excluded left-censored data, but included right-censored data for the time to clearance. We considered three possible distributions of time: exponential, Weibull or log-logistic. The exponential distribution was considered the most parsimonious, as it assumed that the risk of the event occurring (colonization or clearance) was constant over time.

The models potentially included two qualitative explanatory factors in all the analyses – the host and the phylogroup of the focal clone –, and the following quantitative factors: the total density of *E. coli* cells (in CFU/g) and the number of different other clones, of the same phylogroup and total, in the host. We also considered the cell density of the focal clone for the AFT of residence time in the USA dataset, but not for the French dataset, for which this density was not measured. Because of the wide variation in cell densities observed between samples, we used their $\log_{10}$-transformed values for the analyses.

The selection of the best AFT model is presented in the Supporting information. For each response variable and dataset, we generated every possible model including any combination of explanatory factors and time distribution and selected the best model based on Akaike's information criterion [AIC, 44]. We ranked the models according to their ΔAIC, i.e. the difference between their own AIC and the smallest, and kept the set of all models with $\Delta\text{AIC} < 2$. We selected the most parsimonious model among them, unless a likelihood ratio test indicated that a more complex model from this subset, in which the most parsimonious model was nested, explained the data better. We detailed the coefficients of the best model to assess the impact of explanatory factors on time to colonization or clearance (Supporting information).

Correlation of residence time with colonization rates, pathogenicity, and the number of virulence-associated genes

We estimated residence times and colonization rates from the AFT models for both datasets. For a given phylogroup, we used the predicted value for the average host carrying this phylogroup and an average value of all the continuous factors retained in the model selected. The model directly predicted residence times and we computed colonization rates as the inverse of the time to colonization predicted by the model.

Besides, we estimated pathogenicity of each phylogroup on the 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.

In order to assess the correlation between the different variables (residence time, colonization rate, pathogenicity, number of VAG, in total and of each type), we generated for each of them 1 million sets, each comprising a value for each phylogroup, sampled in their respective distributions defined by the estimate and standard error provided by the statistical model. We then computed 1 million Spearman's correlation coefficients between each set of residence time and each set of the other variables. Correlation was considered significantly different from zero if 95% of its distribution was different from zero.

Model description

We developed a continuous time, discrete-state Markov chain model describing the dynamics of *E. coli* clones in a host population. In this model, the discrete states were the strain composition of the hosts, expressed as unordered multisets of the phylogroups of the clones they contained. With G possible phylogroups and N maximum number of clones per host, hosts could be in one of $M = \binom{G+N}{G}$ possible states, which were numbered as follows:

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

The variable of the model was the probability $P_i(t)$ 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:

$$F_k(t) = \sum_{i=1}^{M} V_{i,k} P_i(t) \ , \ \ V_{i,k} = \begin{cases} 0 & \text{if } \ i = 1 \\ m_{i,k}/l_i & \text{if } \ i > 1 \end{cases} \tag{6}$$

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. $V_{i,k}$ corresponded to the proportion of clones of phylogroup k shedded by a host in state i.

The transition between states was defined by the matrix $T(t)$ of size $M \times M$, whose elements were either a clearance rate, noted μ_{k,n,n_k}, or a colonization rate, noted $\lambda_{k,n}(t)$, both defined following an AFT model with an exponential distribution:

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

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 $b < 0$ or as $d > 0$, respectively if it increased the clearance or decreased the colonization of clones as n increased. Stronger competition between clones of the same phylogroup was expressed as $b_s < 0$. The transition matrix $T(t)$ 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 $T(t)$ summed to 0, as the diagonal element $\tau_{i,i}(t) = -\sum_{j \neq i} \tau_{i,j}(t)$. To get transition probabilities between states in continuous time, i.e. accounting for the possibility of multiple transitions over a single day, we computed the matrix exponential $A(t) = e^{T(t)}$ using the package *expm* [66].

We simulated variations in probabilities of the states in discrete time with a time-step of a day, by updating the elements of the transition matrix $T(t)$ using the phylogroup frequencies of the previous day $F_k(t-1)$. While clearance rates were constant over time, colonization rates were updated to account for changes in phylogroup frequency. We ran those simulations until 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^*.

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, 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] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] David M Gordon, Olivier Clermont, Heather Tolley, and Erick Denamur. Assigning Escherichia coli strains to phylogenetic groups: Multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol*, 10(10):2484–2496, 2008. doi: 10.1111/j.1462-2920.2008.01669.x.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] K Totsuka. Studien über Bacterium coli. *Z Hyg Infektionskr*, 45:115–124, 1903.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

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

[65] 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.

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