Persistence-colonization trade-off and niche differentiation enable the coexistence of *E. coli* phylogroups

Thibaut Morel-Journel[a,1], Sonja Lehtinen[b], Olivier Cotto[c], Jonathan Martinson[d], Olivier Clermont[e], Seth Walk[d], Erick Denamur[e,f], and François Blanquart[a]

[a]Center for Interdisciplinary Research in Biology (CIRB), College de France, CNRS, INSERM, Université PSL, Paris, France
[b]Department of Environmental System Science, Institute for Integrative Biology, ETH Zürich, Zürich, Switzerland
[c]PHIM Plant Health Institute, INRAE, Univ Montpellier, CIRAD, Institut Agro, IRD, Montpellier, France
[d]Department of Microbiology and Immunology, Montana State University, Bozeman, MT, USA
[e]IAME, Université Paris Cité and Université Sorbonne Paris Nord, INSERM, Paris, France
[f]AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, Paris, France

Keywords

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

Abstract

Despite the extensive literature on the pathogenicity and virulence of the commensal bacterial species and major pathogen *Escherichia coli*, much less is known about its natural ecological and evolutionary dynamics in populations of healthy hosts. Based on remarkably detailed longitudinal data on the gut microbiota of eight healthy individuals over months to years, we identified a trade-off between colonization and persistence, with major *E. coli* strains (phylogroups) exhibiting a diversity of strategy from strong colonizers persisting for a few days, to poor colonizers persisting for years in the gut. All phylogroups exhibited similar fitness along the trade-off. Moreover, the persistence time was reduced more strongly when a focal strain competed with other strains of the same phylogroups than of different phylogroups, suggesting niche differentiation within the host. To derive the dynamical 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. The trade-off and niche differentiation act together as equalizing and stabilizing mechanisms enabling the stable coexistence of phylogroups over extended periods of time. Strains with the largest number of virulence genes and highest pathogenicity are those persisting for longer in hosts. Our model predicts that a reduction in transmission (e.g., through better hygiene) would not alter the balance between phylogroups, while less disturbance of the microbiome favors persisters, such as those belonging to the pathogenic phylogroup B2.3. More generally, the trade-off between colonization and persistence could play a role in the diversification of other bacterial species of the microbiome.

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 on the order of one million deaths worldwide each year [8, 9, 10]. There is no strict phylogenetic separation between commensal and pathogenic *E. coli* [11], the transition to infection being the result of complex and hardly predictable interactions between several genetic events [12]. Furthermore, genes associated with virulence, i.e. the severity of infections caused, could evolve because of their role in commensalism [13, 14]. Studying the

dynamics of commensal *E. coli* is therefore essential to better understand its dynamics of pathogenicity, i.e. the capacity to cause infections.

As a species, *E. coli* is not only highly diverse, but also very 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 also structured by sequence types [16], which also differ in their characteristics and are themselves congruent with phylogroups [22]. Understanding how this structure has arisen and been maintained is critical to predicting the future evolution of commensal strains of *E. coli*.

More generally, 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 first mechanism is a recombination rate negatively correlated with genetic distance, which can lead to genetically distinct groups even without selection [23, 25]. A second mechanism is ecological niche diversity [26, 27]. Notably, the combination of a specific immune response and a general host immune response favors coexistence among human commensal bacteria, [28]. A third mechanism is the existence of trade-offs between life-history traits, which may allow a species to diversify into distinct strategies [29, 30].

Explaining coexistence between *E. coli* strains requires characterizing their epidemiological dynamics with longitudinal sampling of the same hosts over time. Such studies have been performed with diseased individuals [31, 32], notably to assess the impact of antibiotic use [33, 34], and remain rare in healthy individuals, in partdue to the complexity of setting up such studies. Among studies on healthy individuals, some focus on the special case of initial colonization of infants during their first months of life [35, 36, 37]. The others face the complexity of frequent sampling over a long period and often include a limited number of data points [38, 39].

For this study, we investigated the colonization dynamics of *E. coli* strains belonging to different phylogroups using remarkably detailed longitudinal sampling of the gut microbiome, presented in a previous study by Martinson et al. [40]. Based on the analysis of this dataset, we propose the trade-off between colonization and persistence, combined with niche differentiation between phylogroups, as a novel mechanism facilitating the coexistence of multiple strains among commensal bacteria.

Results

Survival analyses

We studied the turnover dynamics of commensal *E. coli*, using a dataset including an accurate estimation of epidemiological parameters, a longitudinal sampling over several months and fine-grain identification of the clones sampled. Eight healthy individuals were sampled longitudinally over periods ranging from 245 to 849 days [40]. Clones detected during each sampling were identified using 16S rRNA and assigned to phylogroups according to Clermont et al. [17, 18]: A, A.1, B1, B2.1, B2.2, B2.3, C, D, E or F. As a single clone of group C was recorded, it was discarded and this phylogroup was not considered. Phylogroups B1, B2.3, A and F together accounted for around 95% of the *E. coli* cells shedded by volunteers over the entire sampling period (Fig. S1).

Using survival models, we studied the impact of various factors on the frequency with which clones colonize new hosts or are cleared. We mainly present the outcomes of parametric accelerated failure time (AFT) model. Cox models confirm the robustness of the results to the type of statistical model used. Details of both types of models are presented in the Supporting information. The response variables considered are the colonization rate of clones and their persistence time in the gut. We performed a

model selection based on Akaike's information criterion (AIC) [41] and studied the impact of the factors included in the best model.

Colonization rate

The best model describing variations in colonization rate (see Supporting information for model selection) included the following explanatory factors: host identity, phylogroup and number of distinct clones in the host (Table S1). Phylogroup B2.3 had a significantly lower colonization rate than all others (Table S5). Moreover, colonization was more frequent in hosts already harboring a greater number of different clones (95% $CI = [-0.683, -0.271]$, $p < 0.0001$; negative coefficient indicate an increase of the colonization rate, see Supporting information), with a predicted rate multiplied by 1.67 for each additional clone. However, there was no additional impact of the number of clones of the same phylogroup, or of the density of clones in the host. Additional analyses using Cox models with the same factors showed similar relationships, confirming the robustness of these results to the type of survival model used (Supporting information).

Persistence time

The best model describing variations in persistence time (see Supporting information for model selection) included the following explanatory factors: host, phylogroup, number of distinct clones in the host, number of clones of the same phylogroup as the one considered and density of the clone considered. This model highlighted host- and phylogroup-related differences, with phylogroups B2.3, D and F showing the longest persistence times (Table S6). Furthermore, clones persisted longer if they were at higher densities (95%$CI = [0.083, 0.649]$, $p = 0.0113$), but shorter if they were more numerous (95% $CI = [-0.768, -0.097]$, $p = 0.0114$). Importantly, the data showed evidence for niche differentiation between phylogroups: persistence was even shorter if the other clones in the host belonged to the same phylogroup (95% $CI = [-0.493, -0.148]$, $p = 0.0003$). According to the model, the persistence time was divided by 1.56 for each additional clone of another phylogroup, and by 4.93 for each additional clone of the same phylogroup. It was multiplied by 1.83 when the density of the clone considered was multiplied by 10. Again, additional analyses using Cox models with the same factors confirmed the results of the AFT model (Supporting information).

Trade-off between colonization and persistence

The median colonization rate and median persistence time estimated by the two selected AFT models were negatively correlated across phylogroups (Fig. 1). This shows the existence of a trade-off between the colonization and persistence abilities of the phylogroups. Whereas persistence is expected to depend solely on dynamics within the host, colonization is expected to be influenced by the relative frequencies of phylogroups in the host population. To account for this, we corrected the colonization rates predicted by the model, using the relative frequencies of phylogroups observed in our dataset as a proxy for the relative frequencies of phylogroups in the actual host population. To explore the robustness of the relationship to this proxy, we modulated the frequencies of phylogroups by a parameter $u \in [0, 1]$, which describes the observed differences in frequency among phylogroups (Supporting information). Spearman's rank correlation tests showed a significant negative correlation between colonization rate and persistence time for any value of u (Spearman's $\rho \in [-0.600 - 0.404]$, Fig. S2), showing the robustness of this relationship. Thus, phylogroups are distributed along an axis, ranging from strong colonizers with short persistence times, to poor colonizers with long persistence times (Fig. 1). All the envelopes of the estimated values for each phylogroup overlap the $Y = 1/X$ axis (dashed line in Fig. 1), for which colonization balances

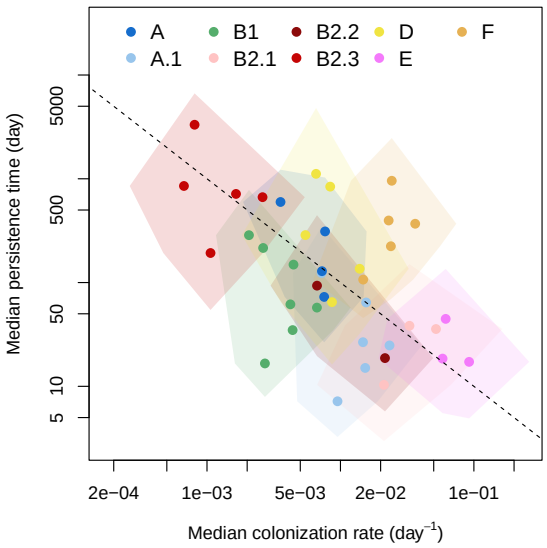

Figure 1: Predicted median persistence time as a function of the predicted median colonization rate, for each combination of host and phylogroup appearing in the data set, for the average number of clones and densities observed in the data. The envelopes include all predicted values ± 2 standard errors for a given phylogroup. The dashed line represents a product of persistence time and colonization rate equal to 1. Results are presented with a correction of colonization rate by phylogroup frequencies with $u = 0.5$.

clearance. Phylogroup F is the most outlying phylogroup, with a fairly high colonization rate relative to its persistence time.

Correlation between persistence and virulence

The trade-off between colonization and persistence has implications for the evolution of bacterial virulence, as the more virulent and pathogenic clones are the persisters. We assessed the differences between phylogroups in terms of their average number of virulence genes carried and their pathogenicity (tendency to cause bloodstream infections), based on a dataset previously analyzed by Marin et al. [42] and Burgaya et al. [43]. First, we show that the median persistence time estimated from the AFT model tends to be greater for phylogroups carrying a greater number of virulence genes on average (Fig 2 A), with two outlier phylogroups (B2.1 and E) accounting for only 4.3% of the data. The positive correlation is also observed when considering only the virulence genes linked to adhesins, and with those linked to iron acquisition and protectins, to a lesser extent. However, we find no correlation when considering those linked to toxins (Supporting information). We also show that the median persistence time also tends to be greater for the most pathogenic phylogroups (Fig 2B), with the outlier group (B2.1) accounting for only 0.9% of the data. Altogether, these consistent results suggest a general positive correlation between the persistence of phylogroups in the gut and their virulence and pathogenicity.

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 *E. coli* clones in a host population. In this model, the discrete states are host gut compositions, i.e. any possible combination of clones, each belonging to one of G phylogroups. States are numbered in increasing order of the number of clones in the host (from 0 to N clones), with $S_1 = \emptyset$, $S_2 = \{1\}$, $S_3 = \{2\}$, $S_{G+1} = \{1, 1\}$, and so on. The variable of the model is the probability $P_i(t)$ that a host is in state S_i at time t.

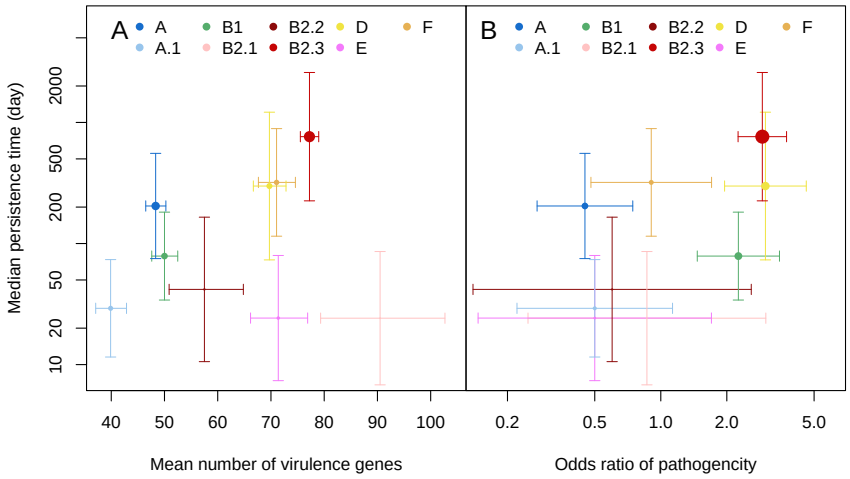

Figure 2: Median persistence time predicted by our AFT model for each phylogroup ±2 standard errors (vertical), as a function of their mean number of virulence genes carried (A) and as a function of their odds ratio of pathogenicity (B), both ±2 standard errors (horizontal). The sizes of the dots correspond to the number of samples considered to estimate the number of virulence genes carried (A) or the pathogenicity (B).

Transitions between states depend on clearance rates μ_{k,n,n_k} and colonization rates $\lambda_{k,n}(t)$, based on those used in the AFT models presented above :

$$\begin{aligned} \mu_{k,n,n_k} &= 1/e^{a_k+b(n-1)+b_s(n_k-1)} \\ \lambda_{k,n}(t) &= \lambda_k^0(t)/e^{c_k+dn} \end{aligned} \quad , \tag{1}$$

with a_k and c_k phylogroup-specific parameters, 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 persistence or colonization corresponding to $b < 0$ and $d < 0$ respectively. Competition generated by niche overlap between clones of the same phylogroup is described by $b_s < 0$.

The impact of phylogroup frequency in the host population on their colonization rate is described by $\lambda_k^0(t)$. Following Levin et al. [44], we model colonization as depending in proportion $(1-q)$ on transmission from other hosts in the population, and in proportion q on influx from a constant reservoir:

$$\lambda_k^0(t) = qF_k^r + (1-q)\,F_k^h(t), \tag{2}$$

with F_k^r the frequency of phylogroup k in the reservoir and $F_k^h(t)$ its frequency in the population. All phylogroups are assumed to be equally represented in the reservoir, so that $F_k^r = 1/G$. In the population, the frequency of phylogroup k is the proportion of clones of phylogroup k in a host, averaged over every host. Based on Eq. 2, $\lambda_{k,n}(t)$ can be expressed as the sum of two rates, corresponding respectively to transmission from the reservoir and transmission within the population:

$$\lambda_{k,n}(t) = \overbrace{qF_k^r e^{-(c_k+dn)}}^{\text{reservoir}} + \overbrace{(1-q)\,F_k^h(t)e^{-(c_k+dn)}}^{\text{within population}}. \tag{3}$$

As $\lambda_{k,n}(t)$ increases with $F_k^h(t)$, the within-population transmission part creates a positive feedback from the phylogroup density on its own colonization rate as in classical epidemiological models.

Phylogroup k can be characterized by a colonization-to-clearance ratio $R_k = e^{a_k-c_k}$, which can also

be interpreted as the basic reproduction number of a clone from this phylogroup. Let us consider a single clone of phylogroup i introduced into an isolated population of H initially uncolonized hosts, such that $q = 0$ and $H \gg 1$. The average frequency of phylogroup i in the population is then $F_i^h(t) = 1/H$ and its expected persistence time is $1/\mu_{i,1,1}$. Assuming that the number of hosts is large enough so that the number of uncolonized hosts $H - 1 \approx H$, the expected number of colonized hosts resulting from this introduction is:

$$H\frac{\lambda_{i,0}(t)}{\mu_{i,1,1}} = H\frac{e^{-(c_i)}/H}{e^{-(a_i+0b_s+0b_o)}} = e^{a_i-c_i} = R_i. \tag{4}$$

Conditions for the coexistence of multiple phylogroups

We computed two indices to assess the role of the observed trade-off on coexistence between strains. Within-host coexistence $E^W(t)$ is the sum of the probabilities that a host is in a state with clones from at least two different phylogroups at time t, with $E^W(t) \in [0, 1]$. Inter-host coexistence $E^I(t)$ is a Shannon index of relative phylogroup frequencies:

$$E^I(t) = -\sum_{k=1}^{G} f_k^h(t)\ln\left(f_k^h(t)\right), \quad f_k^h(t) = \frac{F_k^h(t)}{\sum_{k=1}^{G} F_k^h(t)}, \tag{5}$$

with $E^I(t) \in [0, \ln(G)]$.

For the sake of simplicity, we consider in the following two phylogroups noted i and j ($G = 2$), characterized by their average colonization-to-clearance ratio $\overline{R_{i,j}} = (R_i + R_j)/2$ and their ratio difference $D_{i,j} = R_i - R_j$. We performed discrete-time simulations of our model with a time step of a day and stopped 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 consider this steady state to be the asymptotic state of our system, and note the corresponding intra-host and inter-host coexistence indices E^{W*} and E^{I*}, respectively.

Simulation results show that in the absence of influx from a reservoir and niche differentiation (if $q = 0$ and $b_s = 0$), asymptotic coexistence, i.e. $E^{W*} > 0$ or $E^{I*} > 0$, only occurs for perfectly equally fit strains ($D_{i,j} = 0$) (Fig 3). This means that the trade-off alone allows only structurally unstable coexistence, which is not maintained unless the strains have exactly the same colonization-to-clearance ratios. However, we found that coexistence can be stable for $b_s < 0$, i.e. if niche overlap creates additional clearance. As b_s becomes more negative (greater niche differentiation), the range of values of $D_{i,j}$ for which coexistence is stable widens, but the values of E^{W*} reached also decrease (Fig 3A). 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 coexistence. This is not the case for between-host coexistence, whose maximal value remains the same when b_s varies (Fig 3B). Transmission from a reservoir outside the host population ($q > 0$) also dramatically increases the range of $D_{i,j}$ values for which coexistence is stable (Fig 3C, Fig 3D). This range widens rapidly as q increases, meaning that phylogroup with substantial differences in colonization-to-clearance ratios can coexist even though the impact of the reservoir is limited.

The simulation results are presented here for $\overline{R_{i,j}} = 2$, $b = 0$ and $d = 0$, but we obtain qualitatively similar results for other parameter values (see Supporting information). 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 or low values of d, since these two parameters affect every phylogroup considered equally.

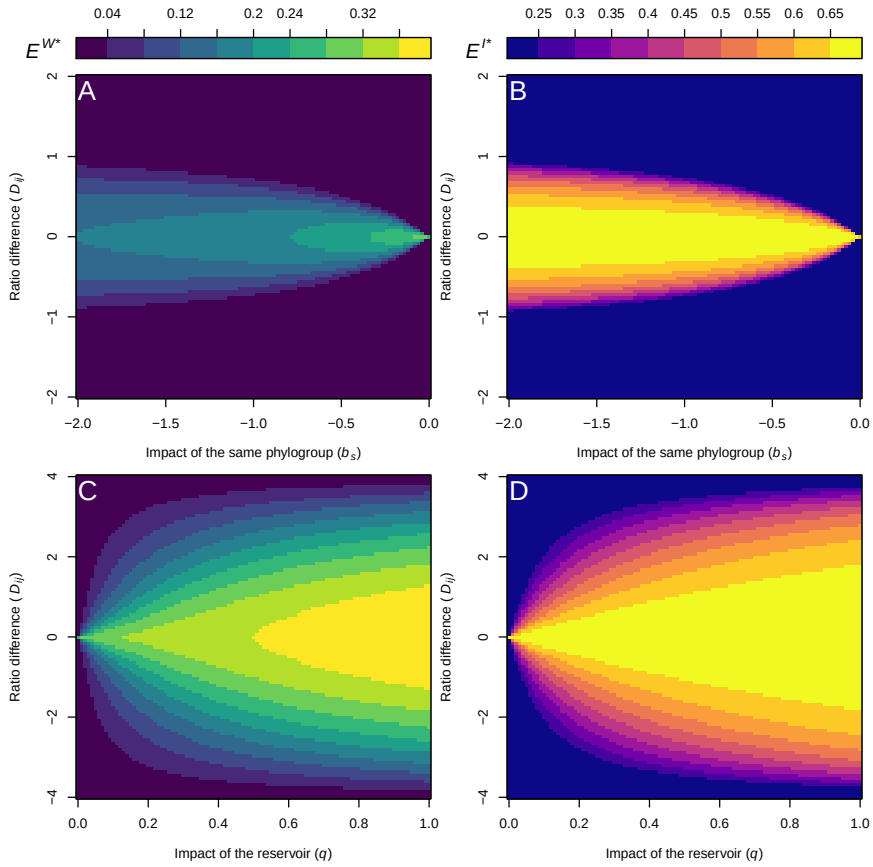

Figure 3: Asymptotic coexistence within hosts E^{W*} (A, C) and inter-hosts E^{I*} (B, D) as a function of the ratio difference $D_{i,j}$ and either the impact of the same phylogroup on clearance b_s or the importance of the reservoir q. For these simulations, $\overline{R_{i,j}} = 2$, $b = 0$, $d = 0$, $E^{W*} \in\ ]0, 1]$ and $E^{I*} \in\ ]0, ln(2)]$.

Impact of changes in colonization or clearance rates on bacterial evolution

Changes in colonization or clearance rates that are not phylogroup-specific, for instance as a result of public health measures, only affect phylogroup relationships if they are additive rather than multiplicative. Indeed, a measure modifying the clearance rate of each clone by a constant factor $\eta > 0$, so that the new rate $\mu^{'}_{k,n,n_k} = \eta\mu_{k,n,n_k}$, also leads to a new ratio $R^{'}_k = R_k/\eta$. Similarly, a multiplicative change in the colonization rate, such that $\lambda^{'}_{k,n}(t) = \eta\lambda_{k,n}(t)$ leads to $R^{'}_k = \eta R_k$. These changes do not alter the hierarchy between phylogroups, i.e. $R^{'}_i > R^{'}_j$ if $R_i > R_j$.

However, a measure adding another source of clearance (e.g. antibiotics) adds another factor θ to the intrinsic clearance rate, so that $\mu^{'}_{k,n,n_k} = \mu_{k,n,n_k} + \theta$. This amounts to multiplying the clearance rate by a phylogroup-specific factor equal to:

$$\mu^{'}_{k,n,n_k} = \theta^m_k \mu_{k,n,n_k}\ ,\ \theta^m_k = \frac{\mu_{k,n,n_k} + \theta}{\mu_{k,n,n_k}}. \tag{6}$$

Hence, the change of colonization-to-clearance ratio is $R^{'}_k = \theta^m_k R_k$. This multiplicative factor θ^m_k is negatively correlated with μ_{k,n,n_k}, meaning that the longest-persisting phylogroup should be the most affected by this additional increase in clearance. Let us assume two clones i and j, such that $R_i > R_j$. Following Eq. 6, there is an additive clearance θ^{eq} that cancels out the difference in ratio between clones, i.e. such that $R^{'}_i = R^{'}_j$:

$$\theta^{eq} = \frac{R_i - R_j}{R_j e^{a_i} - R_i e^{a_j}}. \tag{7}$$

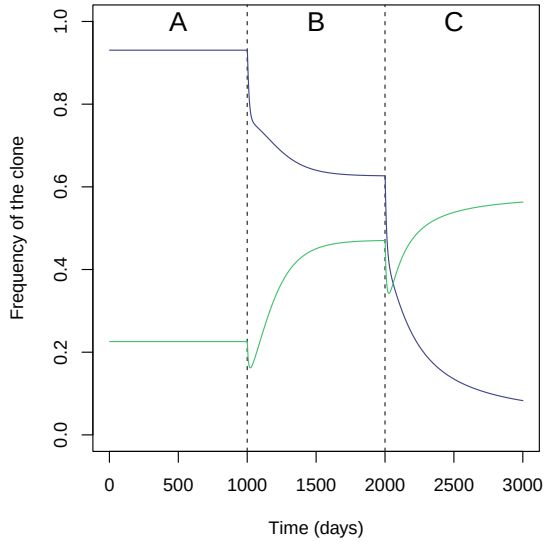

Figure 4: Example of variation in the frequencies of clones i (purple) and j (green) over 3000 simulated days of the model, with $R_i = 3.5$, $R_j = 3$, $\theta_e q = 0.15$, $b = 0$, $b_s = -0.1$ and $d = 0$, . There time-periods of 1000 days each (separated by dashed lines) are defined: (A) no additive 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 j is less affected by the additive clearance, it becomes more frequent than clone i when $\theta > \theta_{eq}$ (C).

It should be noted that $\theta^{eq} > 0$, only if $c_i > c_j$. This means that an additional source of clearance can only negate the advantage of clone i over clone j if the strategy of the latter is more colonization-oriented than that of the former. In that case, the relationship between the two clones is reversed, i.e. $R_j^{'} > R_i^{'}$, if $\theta > \theta^{eq}$, and clone j eventually becomes more abundant than clone i (Fig. 4)

Discussion

Main results

With analysis of finely resolved longitudinal data on *E. coli* colonization dynamics, we explain the stable coexistence between phylogroups of *E. coli* with two mechanisms. The first is the trade-off between the colonization and persistence capacities. Phylogroups are distributed along an axis (represented by the dashed line in Fig 1) ranging from strong colonizers (e.g. E) to long persisters (e.g. B2.3). Interestingly, this distribution is not congruent with the phylogenetic distances between these phylogroups according to Clermont et al. [17]. This trade-off is equalizing *sensu* Chesson [45], as it reduces differences in fitness between phylogroups. In our model, fitness is the colonization-to-clearance ratio R_k: of two clones with different colonization and persistence capacities, the one with the higher value of R_k eventually prevails.

Long-term coexistence also requires a stabilizing mechanism [45]. Our model shows that, while the trade-off alone does not allow two different phylogroups to persist over extended periods of time, niche differentiation could play this role. Indeed, the survival analysis shows a decrease in the persistence time of a clone in the presence of another clone of the same phylogroup in the host. This suggests a greater overlap between the ecological niches of two clones of the same phylogroup, resulting in stronger competition between them. Interestingly, the selected survival model explaining colonization does not include the presence of another clone of the same phylogroup, meaning that the niche overlap only affects clones after the initial colonization of a new host. According to our mathematical model, greater niche differences at the inter-group level allow coexistence between phylogroups with close R_k values, particularly at the host-population level. Transmission from a diverse external reservoir can also maintain

coexistence, by providing an influx of clones of every phylogroup into the host population. Even low reservoir contributions (low q values) can lead to high levels of coexistence.

Our analyses also highlight substantial heterogeneity between hosts, consistent with the absence of 'core resident microbiome' initially reported by Martinson et al. using these data [40]. Our results also suggest that turnover rates vary not only between hosts, but also over time within the same host. Indeed, clearance and colonization are more frequent when a host hosts a greater number of different clones. Thus, hosts experience periods of stability with few clones, alternating with rapid successions of colonization and clearance events, during which the number of clones present temporarily increases. These periods of instability could be triggered by temporal variations in host-related factors affecting the host's suitability for colonization or persistence. 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.

Limitations

Although our dataset provides a very detailed characterization of diversity within the host, the limited number of hosts sampled calls for further, larger-scale studies of the colonization dynamics of commensal *E. coli*. Firstly, the limited size of the data prevented analysis of interactions between specific phylogroups. Secondly, it also prevented an accurate estimation of relative phylogroup frequencies. We showed that the negative correlation between persistence time and colonization rate was rather robust (Supporting information), so that changes in phylogroup frequencies should not invalidate this result. However, we only evaluated cases for which the observed frequencies of phylogroups provide some information about the frequencies in the host population.

For the sake of generality, we modeled colonization as depending on a mixture of human-to-human transmission and transmission from a constant reservoir outside the host population, following [44]. We did not empirically confirm the existence of such a reservoir. A potential candidate would be the environment, which harbors various strains of *E. coli* [46, 47]. This reservoir could also correspond to rarer transmissions from hosts belonging to another population, whose dynamics would be largely independent of phylogroup frequencies in the host population. While our results show that transmission from even a limited reservoir strongly increases coexistence, it is not necessary to explain coexistence between phylogroups. Indeed, niche differentiation inferred from the data already acts as a stabilizing mechanism for coexistence.

Implications

Although this study focuses on commensal *E. coli*, our results also bear implications for the emergence of pathogenic strains. Previous studies have highlighted the group B2, massively represented in our data by B2.3 (over 99%, see Fig. S1) as both persistent and virulent [48, 14]. Our results show a more general positive correlation between the persistence time of a phylogroup and both the number of virulence genes carried and the pathogenicity. This is therefore consistent with the hypothesis that the virulence of *E. coli* as an opportunistic pathogen results from traits linked to its competitive ability as a commensal [14]. Further anaysis at the level of the function suggests that the presence of virulence genes linked to adhesins is especially correlated with persistence, which is consistent with the literature [49].

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 most persistent clones. This disproportionate

impact of antibiotics on the most persistent phylogroup could favor less pathogenic strains, which are better colonizers than persisters. It can also be expected to generate stronger selection for antibiotic resistance among persisters. Yet, unlike the correlation with pathogenicity, there does not seem to be a strong link between persistence and the presence of resistance genes [21, 42]. This differs, for example, from *Streptococcus pneumoniae*, another human commensal bacteria and opportunistic pathogen, for which such a correlation between resistance and persistence has been demonstrated [50].

Conclusion

By re-analyzing longitudinal data on *E. coli* epidemiological dynamics, we discover general principles structuring the diversity of this species: major clones align along a colonization-persistence trade-off equalizing their fitness, and occupy distinct niches in humans, thus stabilizing their long-term coexistence. We also discover a strong correlation between persistence 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

The main dataset used for the survival analyses presented in this study was collected and analyzed by Martinson et al. [40]. It includes longitudinal sampling of healthy patients over periods ranging from 245 to 849 days, with information on the presence and density (in CFU) of each clone at each sampling date. Each clone was assigned to one of nine phylogroups defined according to Clermont et al.[17, 18]: A, A.1, B1, B2.1, B2.2, B2.3, C, D, E or F. As a single clone of group C was recorded, it was discarded and this phylogroup was not considered. Group G was not defined separately by Martinson et al., so it does not appear in the dataset. Different clones of the same phylogroup in the same host were differentiated using GTG5 repetitive element amplification, following Mohapatra & Mazumder [51]. For each clone sampled on each date, we collected the following data : sampling date, host identity, phylogroup of the clone, number of other clones in the host and their phylogroup, total density of *E. coli* in the host and density of the clone considered.

The dataset used to assess the virulence and pathogenicity of phylogroups is a subset of the one collected and analyzed by Burgaya et al. [43]. It includes 704 clones of *E. coli* collected after 2010, with their pathogenic status regarding bloodstream infections (1 for pathogen and 0 for commensal). Each clone was assigned to one of the nine phylogroups listed above, the other clones were discared from the dataset. The dataset also includes the list of virulence genes carried by each of the 248 commensal clones. These genes were distributed into five functional classes associated with virulence factors: toxins (149 genes), adhesin (115 genes), iron acquisition systems (50 genes), protectins (10 genes) and invasins (2 genes). 42 additional miscellaneous genes were included, for a total of 368.

Survival analyses

We used the software *R* [52] and the package *survival* [53] to perform survival analyses on the dataset from Martinson et al. We used accelerated failure time (AFT) models, which describe the main relationship between time to an event – either colonization or clearance – and explanatory factors [54].

The time to colonization by a clone was counted from the last colonization of the host by another clone, and its inverse was defined as the clone's 'colonization rate'. The time to clearance of a clone

was counted from its colonization and defined as its 'persistence time'. Left-censored data were discarded from the analysis, but right-censored data were included. We considered three possible time distributions: exponential, Weibull or log-logistic. The exponential distribution is the most parsimonious, since it assumes that the risk of the event occurring is constant over time. We considered five explanatory factors for these analyses: two qualitative – the host and the phylogroup of the clone – and four quantitative – the densities of the clone (in CFU/g) and of every clones in the host, as well as the numbers of different other clones, of the same phylogroup and total, in the host. We did not include the density of the clone in analyses with colonization as an event, as this clone was not yet present.

The survival analysis had two parts. First, we selected the best model from all possible ones including any combination of explanatory factors and time distribution, based on Akaike's information criterion [AIC, 41]. Given the variations in densities observed between samples, we considered the total and clone's densities as explanatory factors either as is or as their log values. 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. Then, we detailed the coefficients of the best model to assess the impact of explanatory factors on time to colonization or clearance.

Measures of virulence

Using the dataset from Marin et al., we computed the proportion of virulence genes carried by each clone, out of the 368 different genes identified in the dataset, and performed a binomial regression with these proportions as the response variable and the phylogroup of the clones as the explanatory factor to estimate the mean proportion of virulence genes carried by each phylogroup, multiplied by 368 to obtain the mean number. We also computed the mean number of genes in each of the following functional classes of virulence factors: adhesins, toxins, iron acquisition systemis and protectin. As they were only 2 genes associated with invasins, they were not considered separately. Using the dataset from Burgaya et al., we performed a binomial regression with the pathogenic status of clones as the response variable and the phylogroup as the explanatory factor, to get an estimated odds ratio of infection risk of each phylogroup.

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 are the composition of the hosts, expressed as unordered multisets of the phylogroups of the clones they contain. With G possible phylogroups and N maximum number of clones per host, hosts can be in one of $M = \binom{G+N}{G}$ possible states, which are 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 is the probability $P_i(t)$ that a host is in state S_i at time t. The frequency of

phylogroup k in the host population at time t is defined as:

$$F_k^h(t) = \sum_{i=1}^{M} V_{i,k} P_i(t) \ , \ \ V_{i,k} = \begin{cases} 0 & \text{if } \ i = 1 \\ m_{i,k}/l_i & \text{if } \ i > 1 \end{cases} \tag{8}$$

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

The transition between states is defined by the matrix $T(t)$ of size $M \times M$, whose elements are 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:

$$\begin{aligned} \mu_{k,n,n_k} &= 1/e^{a_k + b(n-1) + b_s(n_k - 1)} \\ \lambda_{k,n}(t) &= \lambda_k^0(t)/e^{c_k + dn} \end{aligned} \ , \tag{1}$$

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. $\lambda_k^0(t)$ describes how the frequency of phylogroup k impacts their colonization rate:

$$\lambda_k^0(t) = qF_k^r + (1-q)\, F_k^h(t). \tag{2}$$

Following Levin et al. [44], we consider that colonization depends on proportion $q \in [0, 1]$ from an external reservoir in which phylogroup k is present at a constant frequency qF_k^r, and on proportion $(1 - q)$ from within-population transmission.

The transition matrix $T(t)$ is sparse, as the transition rate $\tau_{i,j}(t) > 0$ only if state j can be reached from state i with a single colonization or clearance event. The rows of $T(t)$ sum 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* [55].

We simulated variations in probabilities of the states in discrete time, by updating the matrix of transition probabilities $A(t)$ at each time-step, i.e. day. To do so, we therefore used $\lambda_{k,n}(t-1)$ to compute the elements of $A(t)$. 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 European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement No. 949208)

Conflicts of interest

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

Author contribution

T.M-J and F.B. designed the project, and conceptualized it with inputs from S.L. J.M. and S.W. produced the first dataset used in this study, while E.D and O. Clermont contributed to the production of the second

one. T.M-J. analyzed the data with the help of O. Cotto and F.B., T.M-J. and F.B. developed the model, with the help of S.L. T.M.-J. led the redaction of the article, with inputs from all the co-authors.

References

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] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] BR Levin, M Lipsitch, V Perrot, S Schrag, R Antia, Lone Simonsen, N Moore Walker, and FM Stewart. The population genetics of antibiotic resistance. *Clin Infect Dis*, 24(Supplement_1):S9–S16, 1997. doi: 10.1093/clinids/24.supplement_1.s9.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

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

[54] 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.

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