

Supplementary material

The following presents supplementary material of the article *Residence-colonization trade-off and niche differentiation enable coexistence of Escherichia coli phylogroups in healthy humans* by Thibaut Morel-Journel, Sonja Lehtinen, Olivier Cotto, Rafika Amia, Sara Dion, Clarisse Figueroa, Jonathan N.V. Martinson, Pascal Ralaimazava, Olivier Clermont, Xavier Duval, Forough L. Nowrouzian, Seth T. Walk, Erick Denamur, François Blanquart.

1. Details of the French dataset
2. Accelerated failure time models
3. Cox models
4. Residence-colonization trade-off among infants
5. Robustness of the results to the number of colonies sampled
6. Analysis of the mathematical model
7. Impact of an external source of *E. coli* clones
8. Supplementary figures
9. Supplementary tables

1. Details of the French dataset

We generated the French dataset presented in the main text of the article as part of the EVE/EvoComBac project. This study has been ethically approved by the institutional review board (Comité de Protection des Personnes Ouest IV - Nantes, board approval number 18/20.3). It was sponsored by Assistance Publique - Hôpitaux de Paris. Consents were obtained in compliance with the French law, with signed informed consent obtained from all participants.

From March 2021 to June 2022, we recruited 50 healthy adults in the Paris area to provide stool samples every two weeks (± 3 days) for a planned total of 10 samples. The inclusion criteria were the following:

- healthy volunteer of both sexes between 18 and 65 years inclusive
- subject considered healthy after a thorough general examination (physical exam and questionnaire)
- subject with normal transit with usually one stool per day
- body mass index between 18.5 and 35 kg/m² inclusive

The non-inclusion criteria were the following:

- subject living in a healthcare institution
- known immunosuppression (HIV, concomitant immunosuppressive treatment, chemotherapy, long-term corticosteroids more than 2 weeks in the last 6 months)
- chronic gastrointestinal diseases (Chronic Ulcerative Colitis, Crohn’s Disease)
- any digestive resection except appendectomy and resection of polyps
- subject with a history of bacteremia

Four of the initially fifty enrolled volunteers dropped out of the study without providing any sample and were discarded. Out of the remaining 46 volunteers, 27 were male, 19 were female and the median age at inclusion was of 23 years old. Some volunteers missed a small number of visits, so that we had stool sample for a total of 430 visits. This small level of attrition, in an ecological study of healthy volunteers without intervention, is unlikely to have biased our results. For each visit, we isolated *E. coli* colonies on Drigalski agar plates. Five randomly chosen colonies were sampled for characterization of within-host diversity. This number was chosen as a compromise between quantifying the diversity of dominant types and limiting the total number of samples to be analyzed. Importantly, we chose the colonies to be sampled independently from their phenotype. We used the Clermont triplex (Clermont et al., 2000) or quadruplex (Clermont et al., 2013) PCR, followed when necessary by phylogroup allele-specific PCRs for C, E and G phylogroups (Clermont et al., 2013, 2019, 2021) to assign each colony to a phylogroup. We identified the bacterial species with MALDI-TOF (mass spectrometry) when the results obtained by Clermont typing were incompatible with the species *E. coli*. To differentiate clones within a same phylogroup, we typed all *E. coli* colonies with multiple-locus variable-number tandem-repeat analysis (MLVA, Caméléna et al., 2019). MLVA gave an electrophoretic pattern with 4 to 8 bands per isolate. Two isolates differing by at least one fragment were considered to be from two distinct clones. This level of discrimination was similar to that of the GTG5 Rep-PCR used for the USA dataset (Mohapatra and Mazumder, 2008). Total *E. coli* density was estimated through serial dilutions. Presence of *E. coli* clones over time in each of the 46 hosts in the French dataset is presented in Fig S11.

2. Accelerated failure time models

Accelerated failure time (AFT) models describe how covariates affect the time until the occurrence of an event, here either the colonization or the clearance of a clone. How covariates accelerate or decelerate the occurrence of an event depends on the acceleration $\gamma = \exp(-\sum_i \alpha_i X_i)$, with X_i the value of the i^{th} explanatory factor considered and α_i the value of the parameter associated with this factor. The acceleration γ correlates positively with the colonization or clearance rate. However, a positive value of α_i indicates a negative impact of the corresponding factor on the rates and a positive impact on the time until the event.

AFT models are parametric, i.e. the time until the event is expected to follow a given distribution. We considered three possible distributions : exponential, Weibull or log-logistic. The exponential distribution is the most parsimonious, as it assumes a constant hazard. Indeed, the other two distributions assume a varying hazard over time, and are characterized by two parameters: the scale, i.e. the inverse of the acceleration factor γ , which is affected by the covariates, and the shape, which is estimated separately.

2.1. Selection of AFT models of residence

We considered the potential impact of the following factors on the residence time, i.e. the time until clearance of each clone:

- the host (*hst*)
- the phylogroup (*grp*)
- the total number of other clones in the host (*ncl*)
- the number of other clones in the host of the same phylogroup as the clone considered (*ngrp*)
- the \log_{10} of the total density of cells in the host ($\log(dns)$)

For the analysis of the USA dataset, we also considered the \log_{10} of the density of cells of the clone considered ($\log(dscl)$), which was precisely known thanks to the large number of colonies typed.

We fitted AFT models with every possible combination of the model for the distribution of time until clearance, and of these explanatory factors. We computed the AIC of each model and order them by increasing ΔAIC , i.e. the difference between the smallest AIC value and that of the model (see Table S1 and S3 for all models with $\Delta AIC < 10$ for the French and USA data, respectively). We selected the most parsimonious model among those with $\Delta AIC < 2$, unless a likelihood ratio tests showed that another model in this subset explained the data better.

For the French data, two nested models had a $\Delta AIC < 2$:

- **model 1:** $Y \sim hst + grp + ncl + ngrp$
- model 2: $Y \sim hst + grp + ncl + ngrp + \log(dns)$

Both models used a log-logistic distribution. We selected model 1 (in bold) because model 2 did not explain the data significantly better (LRT: $\chi^2_{df=1} = 0.650$, P value = 0.420).

For the USA data, four models had a $\Delta AIC < 2$. Three of them used an exponential distribution and were nested within one another:

- model 1: $Y \sim hst + grp + ncl + ngrp$
- **model 2:** $Y \sim hst + grp + ncl + ngrp + \log(dscl)$
- model 3: $Y \sim hst + grp + ncl + ngrp + \log(dns) + \log(dscl)$

The last model used a Weibull distribution:

- model 4: $Y \sim hst + grp + ncl + ngrp + \log(dscl)$

We selected model 2 (in bold) over the two others using an exponential distribution because it explained data significantly better than model 1 (LRT: $\chi^2_{df=1} = 3.891$, P value = 0.049) and as well as model

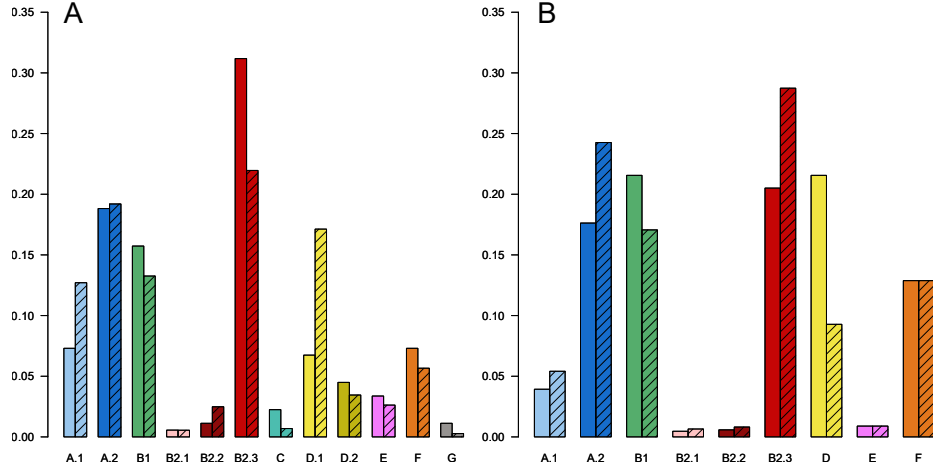


Figure S1: Frequencies of the phylogroups according to the external dataset (clear) and in the dataset itself (hatched), used to correct the colonization rates in the French (A) and in the USA data (B).

3 (LRT: $\chi^2_{df=1} = 1.013$, P value = 0.314). We also selected model 2 over model 4 because, although they included the same explanatory factors, the former was more parsimonious as it assumed a constant clearance hazard.

Thus, the models best explaining the residence time of *E. coli* in France and USA were very similar. They included the same explanatory factors, except for the addition of the density of the clone considered in the model for the USA dataset, a factor that was not available for the French data.

2.2. Correction of the colonization rate by the phylogroup frequency

In addition to the explanatory factors taken into account in the AFT models (see below), we also expected colonization to be impacted by the relative frequencies of the phylogroups in the population. Indeed, more prevalent phylogroups were likely shed in greater quantities, and therefore contributed more to transmission. We assumed as a null model that phylogroups were shed in proportion to their frequency in the population. Accordingly, we corrected the observed colonization rates by the frequency of the phylogroup in the host population. This was equivalent to multiplying observed colonization times by phylogroup frequency.

We assessed phylogroup frequencies in the host population based on two types of data relevant to the country considered: external data and the dataset itself (Fig S1). For the French dataset, we used cross-sectional external data on the carriage of *E. coli* in healthy hosts of the Paris area, previously analyzed by Burgaya et al. (2023). This dataset is presented in the main text of the article, as it was also used to assess the pathogenicity of the phylogroups and their number of virulence associated genes (see the Material and Methods section). The phylogroups of these 359 clones were defined in the same way as those considered for the French dataset. For the USA dataset, we considered six studies of commensal *E. coli* carriage conducted in the USA (Zhang et al., 2002, Sannes et al., 2004, Johnson et al., 2005, 2007, Hannah et al., 2009, Logue et al., 2012). While they all reported the frequencies of the main phylogroups of *E. coli* (A, B1, B2 and D), they did not necessarily included the minor phylogroups (E and F), which represented 5.1% of the clones in the USA dataset. We therefore assumed that these frequencies corresponded to the ones observed in the USA dataset itself. Besides, these external data did not make the distinction between phylogroups A.1 and A.2 (all noted ‘A’) or between B2.1, B2.2 and B2.3 (all noted ‘B2’). We inferred the relative frequencies of these subgroups based on the USA dataset. For instance, as A.1 and A.2 represented respectively 42% and 58% of the A identified in our dataset, we assigned respectively

42% and 58% of the clones of phylogroup A observed in the external data to A.1 and A.2.

Although the frequencies observed in the external data were comparable to the ones observed in both our datasets (Fig S1), some phylogroups differed in their frequency between the two. To assess whether these differences impacted the negative correlation between residence time and colonization rate observed (see the Results section of the main text of the article), we performed the same analyses using a correction based on the phylogroup frequencies in the dataset themselves. The results were very similar the one presented in the main text, confirming the robustness of this trade-off for the French (95%CI = $[-0.618, -0.832]$, P value < 0.001) and USA datasets (95%CI = $[-0.900, -0.583]$, P value < 0.001).

2.3. Selection of AFT models of colonization

We used the same model selection method as for the AFT models of residence. We considered the potential impact of the following factors on the time until colonization by each clone:

- the host (hst)
- the phylogroup (grp)
- the total number of other clones in the host (ncl)
- the number of other clones in the host of the same phylogroup as the clone considered ($ngrp$)
- the \log_{10} of the total density of cells in the host ($\log(dns)$)

We fitted AFT models with every possible combination of the model for the distribution of time until colonization and of these explanatory factors. We computed the AIC of each model and order them by increasing ΔAIC , i.e. the difference between the smallest AIC value and that of the model (see Table S6 and 8 for all models with $\Delta AIC < 10$ for the French and USA data, respectively). We selected the most parsimonious model among those with $\Delta AIC < 2$, unless a likelihood ratio tests showed that another model in this subset explained the data better.

For the French data, three nested models had a $\Delta AIC < 2$:

- **model 1:** $Y \sim hst + grp + ncl$
- model 2: $Y \sim hst + grp + ncl + ngrp$
- model 3: $Y \sim hst + grp + ncl + \log(dns)$

All of these models used the Weibull distribution. We selected model 1 (in bold) because neither model 2 (LRT: $\chi^2_{df=1} = 0.721$, P value = 0.396) nor model 3 (LRT: $\chi^2_{df=1} = 0.456$, P value = 0.499) explained the data significantly better.

For the USA data, four nested models had a $\Delta AIC < 2$:

- **model 1:** $Y \sim hst + grp + ncl$
- model 2: $Y \sim hst + grp + ncl + ngrp$
- model 3: $Y \sim hst + grp + ncl + \log(dns)$
- model 4: $Y \sim hst + grp + ncl + ngrp + \log(dns)$

All of these models used the Weibull distribution. We selected model 1 (in bold) because neither model 2 (LRT: $\chi^2_{df=1} = 1.815$, P value = 0.178), model 3 (LRT: $\chi^2_{df=1} = 0.650$, P value = 0.420) nor model 4 (LRT: $\chi^2_{df=2} = 2.534$, P value = 0.282) explained the data significantly better. As for the models of residence, the models best explaining the colonization rates in France and the USA were consistent. Indeed, the explanatory factors included were the same for both datasets.

3. Cox models

To confirm the results of the survival analyses using the AFT models (see the Results section of the main text of the article), we compared them to those of semi-parametric Cox models including the same explanatory factors. These models compute hazard ratios indicating the relative risk of events – colonization or clearance – occurring depending on the factors considered. Unlike AFT models, Cox models make no assumptions about the shape of the hazard over time, but assume proportional hazards between strata. In those models, a factor increasing the hazard decreases the time until the event. Therefore, positive coefficients in one type of model corresponds to negative coefficients in the other.

The host factor systematically violated the proportional hazard hypothesis, for residence (French dataset: $\chi^2_{fd=45} = 123.3161$, $\text{textit{P value}} < 0.001$; USA dataset: $\chi^2_{fd=7} = 24.75$, $P \text{ value} = 0.001$) and for colonization (French dataset: $\chi^2_{fd=7} = 139.1213$, $P \text{ value} < 0.001$; USA dataset: $\chi^2_{fd=7} = 51.70$, $P \text{ value} < 0.001$). Therefore, we used Cox models stratified by hosts. Otherwise, we included the same explanatory factors as those used in the corresponding best AFT models.

The Cox models provided results consistent with the AFT models (Table S11-S14). The sign and significance of the coefficients associated with quantitative factors were consistent. Although the significance of the coefficients associated to the phylogroups was sometimes different in the Cox model, their sign was also always consistent with that of the AFT models. These complementary analyses therefore confirm our initial observations.

4. Residence-colonization trade-off among infants

We assessed whether the trade-off between residence and colonization identified among adults (see the Results section of the main text of the article) was also observed during the initial colonization of the gut of infants. To do so, we used an additional longitudinal dataset, previously analyzed by Ostblom et al. (2011). It included longitudinal samples in 130 infants at fixed times since birth (after 3 days, 1, 2 and 4 weeks, 2, 6 and 12 months). The 273 clones of *E. coli* identified in these samples were assigned to one of the four main *E. coli* phylogroups (A, B1, B2 and D).

We used accelerated failure time (AFT) models to analyze these data, using the same method, explanatory factors and time distributions as described above (see Supplementary material 2). To correct the observed colonization rates, we used the phylogroup frequencies in the dataset itself. No external study provided the frequency of *E. coli* phylogroups in healthy infants in Sweden, as the other published studies were based on the same cohort (Nowrouzian et al., 2005, Karami et al., 2007)).

We used the same model selection method as presented in Supplementary material 2, based on AIC (see Table S5 and S10 for all models with $\Delta AIC < 10$ for residence and colonization, respectively). Four nested AFT models explaining the residence time had a $\Delta AIC < 2$:

- **model 1:** $Y \sim grp + ncl$
- model 2: $Y \sim grp + ncl + \log(dns)$
- model 3: $Y \sim grp + ncl + \log(dncl)$

All these models used a Weibull distribution. We selected model 1 (highlighted in bold) because neither model 2 (LRT: $\chi^2_{df=1} = 0.065$, P value = 0.798), model 3 (LRT: $\chi^2_{df=1} = 0.190$, P value = 0.663) nor model 4 (LRT: $\chi^2_{df=1} = 0.135$, P value = 0.714) explained the data significantly better. The number of clones in the same phylogroup was not a factor in the model that best explained residence time for this data set, in contrast to the best AFT models for French and American data.

Three nested AFT models explaining the colonization rate had a $\Delta AIC < 2$:

- **model 1:** $Y \sim grp + ncl$
- model 2: $Y \sim grp + ncl + ngrp$

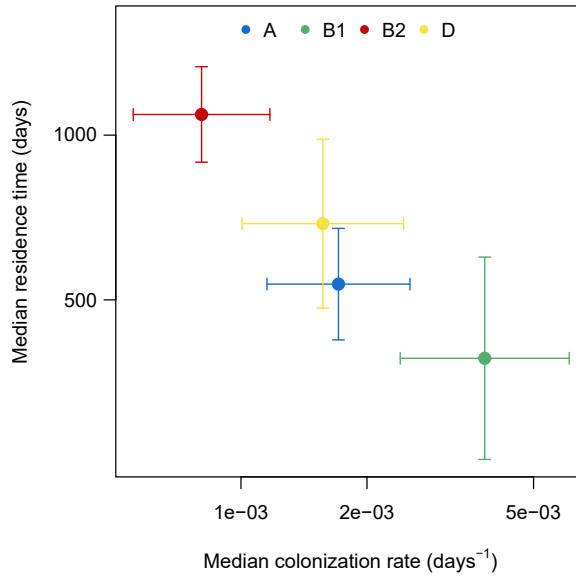


Figure S2: Median residence time as a function of the median colonization rate, as estimated for each phylogroup in the infant dataset, for an average number of clones. Dots represent estimates and bars their 95% confidence intervals.

– model 3: $Y \sim grp + ncl + \log(dns)$

These models all used the Weibull distribution. Again, we selected model 1 because neither model 2 (LRT: $\chi^2_{df=1} = 0.085$, P value = 0.771) nor model 3 (LRT: $\chi^2_{df=1} = 0.263$, P value = 0.608) explained the data significantly better. The host factor was included in neither models best explaining residence or colonization. This was likely due to the very large number of infants (130) which reduced considerably the parsimony of models including this explanatory factor.

The selected models provided estimates and standard errors for the residence times and colonization rates (Fig S2). We observed a trade-off between the colonization and residence abilities of *E. coli* phylogroups during the initial colonization of the gut of infants during their first year. The negative correlation was significant (parametric bootstrapping; Spearman's $\rho = -0.969$, 95%CI = $[-0.996, -0.496]$, P value = 0.001).

5. Robustness of the results to the number of colonies sampled

While the French dataset encompassed a larger number of subjects, the number of colonies used for each sample (five) was lower than that of the USA dataset, which was of the order of 95 (Martinson et al., 2019). We tested how this difference affected our ability to detect the intra-host diversity of clones within the samples, and the conclusions of our AFT models.

We compared the diversity observed in the two datasets. Diversity was computed as the number of distinct clones per sample. The diversity was comparable, and slightly higher in the French population, in spite of less exhaustive sampling. Diversity was 1.74 clones per sample in France, compared with 1.56 in the USA (Fig S3A). Geographical differences in clone diversity could explain the slightly higher diversity in France (Escobar-Páramo et al., 2004, Skurnik et al., 2008). Yet, the distribution of the number of clones dropped rapidly in both datasets, indicating that colonization by more than five clones was rare.

We assessed how diversity could be reduced by the typing of a limited number of colonies. We sub-sampled the USA dataset by randomly selecting five colonies by sample without replacement. Therefore, the probability that a clone appeared in a sub-sampled dataset was proportional to the fraction of colonies of this clone in the sample. Sub-sampling reduced the intra-host diversity to 1.24 clone per sample (Fig S3B). This confirms that the diversity of *E. coli* in France is even larger than suggested by Fig S3A.

We then investigated how this sub-sampling could alter our results on niche differentiation and the colonization-residence trade-off. We performed survival analyses for residence and colonization on the sub-sampled datasets, using the same time distribution and explanatory factors as for the analysis of the complete data. The impact of the number of clones of the same phylogroup was only significant, i.e. with a P value < 0.05 , in 10.3% of the AFT models, presumably due to the more limited diversity caused by sub-sampling. However, there was no evidence for a bias in the estimates of this parameter when diversity was reduced. Indeed, the model estimate for the complete data (coef = -1.108) was in the 44th percentile of the distribution of estimates for the sub-sampled data (Fig S4).

We also computed the predicted median residence time and colonization rate of each phylogroup from the AFT models on all the subsets, and measured their correlation. Despite the sub-sampling, the estimates were overall similar to those based on the complete USA dataset (Fig S5). Main differences

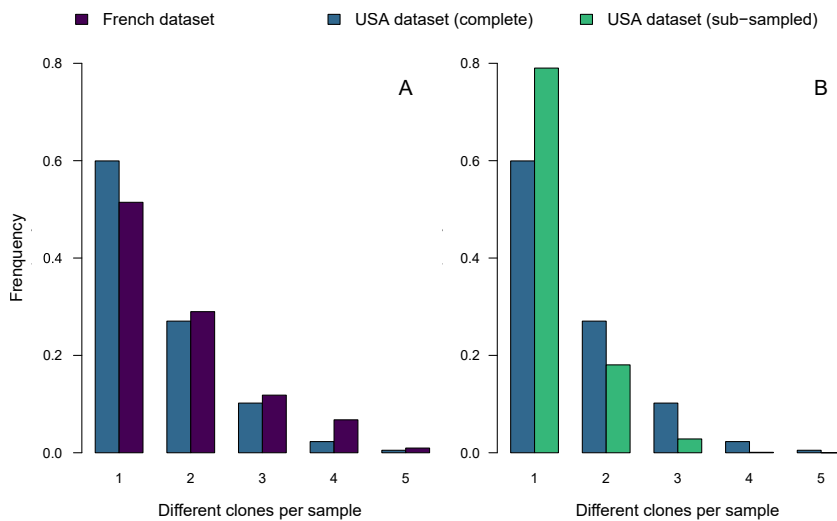


Figure S3: Distribution of the number of different clones per sample in the French dataset (purple), USA dataset (blue) and the subsets of the USA dataset created by randomly selecting five colonies by sample (green). A compares the diversity between the two datasets. B compares the diversity in the USA dataset with and without sub-sampling.

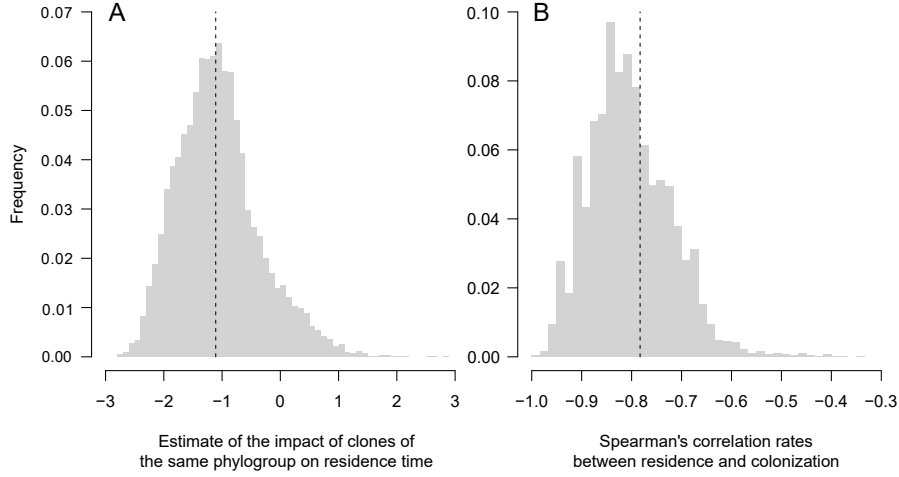


Figure S4: Distribution of the estimates of the impact of the number of clones of the same phylogroups on the residence time of the clones (A) and of the Spearman's correlation between residence time and colonization rate (B), for the sub-sampled USA dataset (grey) and for the complete dataset (dashed line).

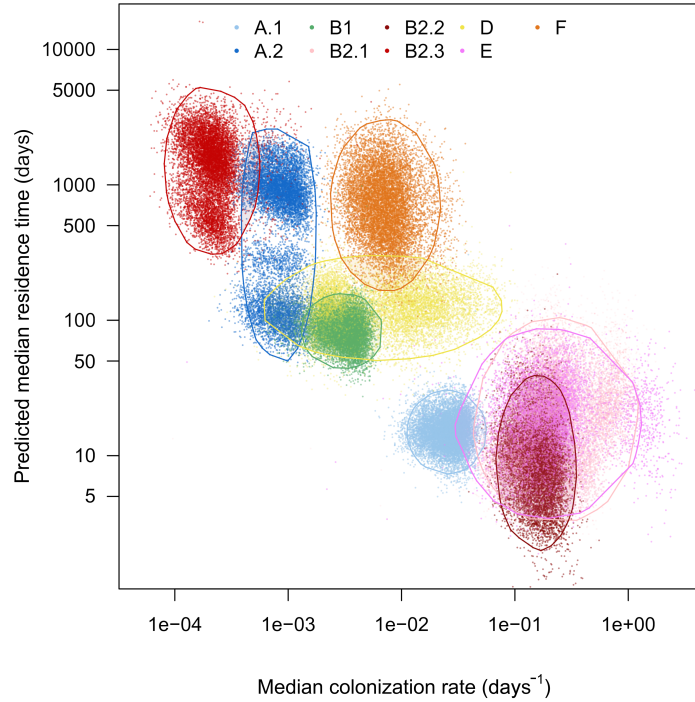


Figure S5: Median residence time as a function of the median colonization rate for each phylogroup according to analyses of the reconstructed datasets. Dots represent individual estimates from the analysis of each dataset. Polygons encompass 95% of the estimates for a given phylogroup.

concerned the phylogroup A and D, whose estimated survival were respectively lower and higher. However, the correlation between colonization and residence remained largely negative for the reconstructed datasets, with 95% of the distribution of Spearman's coefficients between -0.93 and -0.63 . Once again, there was no strong bias induced by sub-sampling, as the correlation rate for the complete data (Spearman's $\rho = -0.733$) was in the 64th percentile of the distribution.

6. Analysis of the mathematical model

This section presents the analysis of the mathematical model describing *E. coli* colonization dynamics presented in the main text. Firstly, we present the model, which describes how the colonization status of hosts changes by the events of colonization and clearance. Secondly, we examine what happens when a single bacterial phylogroup is present in the host population. We show how the invasion dynamics of the phylogroup in an ‘empty’ (non-colonized) host population depends on the colonization-to-clearance ratio R_k , which is also the basic reproduction number in epidemiology. We derive the equilibrium endemic distribution of the number of clones of this phylogroup, and the dependence on competition parameters. Thirdly, we consider a case with two phylogroups. We show that under some conditions on competition (symmetric competition among phylogroups, no niche differentiation), a new phylogroup can invade the population at endemic equilibrium if and only if its colonization-to-clearance ratio R_k is greater than that of the resident strain. Finally, we show how changes in transmission or clearance can alter the competitive hierarchy between phylogroups.

6.1. Presentation of the model

In our model, a clone of phylogroup k in a host with n colonizing strains is characterized by a clearance rate μ_{k,n,n_k} , such that:

$$\mu_{k,n,n_k} = 1/e^{a_k+b(n-1)+b_s(n_k-1)} = e^{-a_k} e^{-b(n-1)} e^{-b_s(n_k-1)}. \quad (1)$$

When there is no niche differentiation creating additional clearance of clones of the same phylogroup, i.e. $b_s = 0$, the clearance rate can be expressed as:

$$\mu_{k,n,n_k} = e^{-a_k} e^{-b(n-1)}. \quad (2)$$

We expect the coefficient b to be negative if more clones lead to a higher clearance rate, with e^{-a_k} corresponding to the clearance rate in hosts with a single clone.

The clone is also characterized by a colonization rate $\lambda_{k,n}(t)$, such that:

$$\lambda_{k,n}(t) = F_k(t)/e^{c_k+dn} = F_k(t)e^{-c_k}e^{-dn}. \quad (3)$$

We expect coefficient d to be positive if more clones lead to lower colonization rates. In contrast to the clearance rate, the colonization rate of phylogroup k depends on its frequency $F_k(t)$ in the host population, defined as:

$$F_k(t) = \sum_{j=1}^M V_{j,k} P_j(t), \quad (4)$$

with $P_j(t)$ the frequency of state j , and $V_{j,k}$ the frequency of phylogroup k in host state j . We consider here that the phylogroup is transmitted proportionally to its frequency in each host state, i.e. $V_{j,k} = m_{j,k}/l_j$ with $m_{j,k}$ the number of clones of phylogroup k and l_j the total number of clones in state j . As a convention, $V_{1,k} = 0$, as state 1 corresponds to an empty host.

6.2. Model with a single phylogroup

Let us consider a single phylogroup k ($G = 1$), such that the frequency of hosts colonized by n clones of this phylogroup is noted $X_{k,n}(t)$. All frequencies sum to 1, i.e. $\sum_{n=0}^{\infty} X_{k,n}(t) = 1$. The dynamics of the

frequencies of hosts can be expressed as:

$$\begin{aligned}\dot{X}_{k,0} &= -\lambda_{k,0}(t)X_{k,0} + \mu_{k,1,1}X_{k,1} \\ \dot{X}_{k,n} &= \lambda_{k,n-1}(t)X_{k,n-1} - \lambda_{k,n}(t)X_{k,n} + \mu_{k,n+1,n+1}X_{k,n+1} - \mu_{k,n,n}X_{k,n} \quad \forall n \geq 1\end{aligned}\tag{5}$$

In the first equation, the first term represents colonization of empty hosts, and the second term clearance from hosts with one clone. In the second equation, the first and second terms respectively represent colonization of hosts with $n-1$ and n clones, the third and fourth terms respectively represent clearance of hosts with $n-1$ and n clones. The time dependency of the variables $X_{k,n}$ is omitted for clarity. Replacing the expressions for $\mu_{k,n,n}$ and $\lambda_{k,n}(t)$ according Eq 2 and Eq 3 respectively give:

$$\begin{aligned}\dot{X}_{k,0} &= -F_k(t)e^{-c_k}e^{-d \times 0}X_{k,0} + e^{-a_k}e^{-b \times 0}X_{k,1} \\ \dot{X}_{k,n} &= F_k(t)e^{-c_k}e^{-d(n-1)}X_{k,n-1} - F_k(t)e^{-c_k}e^{-dn}X_{k,n} + e^{-a_k}e^{-bn}X_{k,n+1} \\ &\quad - e^{-a_k}e^{-b(n-1)}X_{k,n} \quad \forall n \geq 1\end{aligned}\tag{6}$$

We can rescale time by e^{-a_k} without loss of generality, such that the mean residence time of the phylogroup k in a host colonized by a single clone is 1. The dynamics described in Eq 6 then becomes:

$$\begin{aligned}\dot{X}_{k,0} &= -F_k(t)\frac{e^{-c_k}}{e^{-a_k}}e^{-d \times 0}X_{k,0} + e^{-b \times 0}X_{k,1} \\ \dot{X}_{k,n} &= F_k(t)\frac{e^{-c_k}}{e^{-a_k}}e^{-d(n-1)}X_{k,n-1} - F_k(t)\frac{e^{-c_k}}{e^{-a_k}}e^{-dn}X_{k,n} + e^{-bn}X_{k,n+1} - e^{-b(n-1)}X_{k,n} \quad \forall n \geq 1\end{aligned}\tag{7}$$

Defining the colonization-to-clearance ratio, i.e. the basic reproduction number of phylogroup k , $R_k := e^{a_k - c_k}$, Eq 7 gives:

$$\begin{aligned}\dot{X}_{k,0} &= -F_k(t)R_k e^{-d \times 0}X_{k,0} + e^{-b \times 0}X_{k,1} \\ \dot{X}_{k,n} &= F_k(t)R_k e^{-d(n-1)}X_{k,n-1} - F_k(t)R_k e^{-dn}X_{k,n} + e^{-bn}X_{k,n+1} - e^{-b(n-1)}X_{k,n} \quad \forall n \geq 1\end{aligned}\tag{8}$$

6.2.1 Invasion dynamics of the phylogroup when rare depends on the colonization-to-clearance ratio R_k

Here we show that the dynamics of a newly introduced phylogroup in the host population depends on the colonization-to-clearance ratio R_k . Let us consider a case where phylogroup k is the only one present in the host population at low prevalence, i.e. $X_{k,0}$ is close to 1, $X_{k,1}$ is close to 0 and $X_{k,n} = 0$ for $n > 1$. As above, the time-dependency of the variables $X_{k,n}$ is omitted for clarity. In this case, F_k is also close to 0. More precisely, $F_k(t) = X_{k,1} = 1 - X_{k,0}$ under our assumptions for transmission. The system then reduces to the first order to:

$$\begin{aligned}\dot{X}_{k,0} &= -F_k(t)R_k e^{-d \times 0}X_{k,0} + e^{-b \times 0}X_{k,1} \\ \dot{X}_{k,1} &= F_k(t)R_k X_{k,0} - X_{k,1} \quad \forall n \geq 1\end{aligned}\tag{9}$$

Then, phylogroup k increases in frequency if and only if $\dot{X}_{k,1} > 0$, meaning that:

$$F_k(t)R_k X_{k,0} - X_{k,1} = X_{k,1}R_k X_{k,0} - X_{k,1} \approx X_{k,1}(R_k - 1) > 0.\tag{10}$$

This condition is equivalent to $R_k > 1$. Therefore, the ability of phylogroup k to invade a population is defined by its colonization-to-clearance ratio R_k .

6.2.2 Endemic equilibrium solutions

Once a phylogroup is introduced and increases in prevalence, it will reach a non-zero prevalence with an equilibrium distribution of the multiplicity of colonization. Here we find the equilibrium solution for the

frequency of each multiplicity of colonization when $R_k > 1$. The equilibrium value for $X_{k,1}$ verifies:

$$\hat{X}_{k,1} = F_k(t) R_k \hat{X}_{k,0}. \quad (11)$$

The next value can be found by recursion, solving for $\dot{X}_{k,n} = 0$:

$$\hat{X}_{k,n+1} = F_k(t) R_k \left(e^{(b-d)n} \hat{X}_{k,n} - e^{b \cdot n - d \cdot (n-1)} \hat{X}_{k,n-1} \right) + e^b \hat{X}_{k,n}. \quad (12)$$

For $n = 1$, Eq 12 gives:

$$\hat{X}_{k,2} = F_k(t) R_k \left(e^{(b-d)} \hat{X}_{k,1} - e^b \hat{X}_{k,0} \right) + e^b \hat{X}_{k,1}. \quad (13)$$

Replacing $\hat{X}_{k,1}$ according to Eq 11 gives:

$$\hat{X}_{k,2} = F_k(t) R_k \left(e^{(b-d)} F_k(t) R_k \hat{X}_{k,0} - e^b \hat{X}_{k,0} \right) + e^b F_k(t) R_k \hat{X}_{k,0} \quad (14)$$

Using the ansatz $\hat{X}_{k,n} = a_n [F_k(t) R_k]^n \hat{X}_{k,0}$ (with $a_1 = 1$), replacing in Eq 12 gives:

$$\begin{aligned} \hat{X}_{k,n+1} &= F_k(t) R_k \left(e^{(b-d)n} a_n [F_k(t) R_k]^n \hat{X}_{k,0} - e^{b \cdot n - d \cdot (n-1)} a_{n-1} [F_k(t) R_k]^{n-1} \hat{X}_{k,0} \right) + \\ &\quad e^b a_n [F_k(t) R_k]^n \hat{X}_{k,0} \\ \hat{X}_{k,n+1} &= \left(e^{(b-d)n} a_n [F_k(t) R_k]^{n+1} \hat{X}_{k,0} \right) + \left(e^b a_n - e^{b \cdot n - d \cdot (n-1)} a_{n-1} \right) [F_k(t) R_k]^n \hat{X}_{k,0} \end{aligned} \quad (15)$$

From the first term, a_{n+1} must satisfy the following recursion:

$$a_{n+1} = e^{(b-d)n} a_n. \quad (16)$$

In the second term, replacing a_n by $e^{(b-d)(n-1)} a_{n-1}$ gives:

$$\left[e^b e^{(b-d)(n-1)} - e^{b \cdot n - d \cdot (n-1)} \right] a_{n-1}. \quad (17)$$

The coefficient in brackets cancels, so that the solution is of the form:

$$\hat{X}_{k,n} = a_n [F_k(t) R_k]^n \hat{X}_{k,0}. \quad (18)$$

The series is therefore as follows:

$$\begin{aligned} a_1 &= 1 \\ a_2 &= e^{(b-d)} \\ a_3 &= e^{2(b-d)} e^{(b-d)} \\ a_4 &= e^{3(b-d)} e^{2(b-d)} e^{(b-d)} \\ a_n &= e^{(b-d) \sum_{i=1}^{n-1} i} = e^{(b-d) \sum_{i=1}^{n-1} i} = e^{(b-d)n(n-1)/2} \end{aligned} \quad (19)$$

Therefore, the general solution for $\hat{X}_{k,n}$ is as follows:

$$\hat{X}_{k,n} = e^{(\mathbf{b}-\mathbf{d}) \frac{n(n-1)}{2}} [F_k(t) R_k]^n \hat{X}_{k,0} \quad (20)$$

The frequency of colonized hosts decreases for large enough values of n if $(b-d) < 0$ and also decreases for any n if $F_k(t) R_k < 1$.

The phylogroup frequency $F_k(t)$ is a linear combination of $\hat{X}_{k,0}$. Under our assumption for transmis-

sion, $F_k(t) = 1 - \hat{X}_{k,0}$ as host can only carry clones of phylogroup k . Eq 20 therefore becomes:

$$\hat{X}_{k,n} = e^{(b-d)\frac{n(n-1)}{2}} R_k^n (1 - \hat{X}_{k,0})^n \hat{X}_{k,0} \quad (21)$$

Solving for $\hat{X}_{k,0}$ with the constraint that all host frequencies must sum to 1 to obtain an explicit expression, Eq 21 gives:

$$\sum_{n=0}^{\infty} \hat{X}_{k,n} = 1 = \hat{X}_{k,0} \sum_{n=0}^{\infty} e^{(b-d)\frac{n(n-1)}{2}} R_k^n (1 - \hat{X}_{k,0})^n \quad (22)$$

A trivial solution is $\hat{X}_{k,0} = 1$, i.e. every clone is cleared and the phylogroup does not persist. The non-trivial solution in $[0, 1)$ does not have an explicit solution, as the series in $\sum_{n=1}^{\infty} e^{-n^2} A^n$ converges but does not have an explicit solution. For $\hat{X}_{k,0}$ close to 0, the right-hand side of Eq 22 is approximately linear in $X_{k,0}$ with a positive slope of value $\sum_{n=0}^{\infty} e^{(b-d)\frac{n(n-1)}{2}} R_k^n$. For $\hat{X}_{k,0}$ close to 1, the series is approximately $1 + (R_k - 1)(1 - \hat{X}_{k,0})$ and decreases linearly with $\hat{X}_{k,0}$. If there is only one value of $\hat{X}_{k,0} \in [0, 1)$ for which the series equals 1, then the series is an increasing function of $\hat{X}_{k,0}$ at this solution. Therefore, $\hat{X}_{k,0}$ is a decreasing function of R_k .

6.3. Model with multiple phylogroups

We now investigate what happens when an invading phylogroup is introduced in the population of the resident phylogroup at equilibrium. In the following, we show that R_k is the fitness quantity determining whether the invading phylogroup will rise in frequency or not. Let us consider a ‘resident’ phylogroup k established at endemic equilibrium and an ‘invading’ phylogroup v at low density in the host population. The system governing the dynamics of the phylogroups can be described as:

$$\begin{aligned} \dot{X}_{k,n} = & F_k(t) R_k e^{-d(n-1)} X_{k,n-1} - \left(R_k F_k(t) + \underbrace{R_v F_v(t)}_{\text{colonization by } v} \right) e^{-d n} X_{k,n} \\ & + e^{-b n} \left(X_{k,n+1} + \underbrace{\frac{Y_{v,n+1}}{n+1}}_{\text{clearance of } v} \right) - e^{-b(n-1)} X_{k,n} \quad \forall n \geq 1 \\ \dot{Y}_{v,n} = & e^{-d(n-1)} \left(\underbrace{F_k(t) R_k Y_{v,n-1}}_{\text{colonization by } k} + \underbrace{F_v(t) R_v X_{k,n-1}}_{\text{colonization by } v} \right) - \underbrace{F_k(t) R_k e^{-d n} Y_{v,n}}_{\text{colonization by } k} \\ & + \underbrace{e^{-b n} \frac{n}{n+1} Y_{v,n+1} - e^{-b(n-1)} Y_{v,n}}_{\text{clearance of } k} \quad \forall n \geq 1 \end{aligned} \quad (23)$$

with $Y_{v,n}$ the density of hosts carrying one clone of phylogroup v and $(n-1)$ clones of phylogroup k $\forall n \geq 1$, R_v the reproduction number of phylogroup v (Fig S6) and $F_v(t)$ is the frequency of phylogroup v .

The first equation differs from the previous equation describing a monomorphic population by two terms. First, the colonization by the phylogroup v reduces the frequency of hosts colonized with n strains of phylogroup k at a rate proportional to $R_v F_v(t)$. Second, hosts of type $(v, n+1)$ become of type (k, n) when losing their clone of phylogroup v . In the second equation, hosts of type $(v, n-1)$ or of type $(k, n-1)$ become of type (v, n) when colonized by a clone of phylogroup k and phylogroup v , respectively. Furthermore, hosts of type $(v, n+1)$ also become of type (v, n) with the clearance of a clone of phylogroup k . Here, we assume that colonization by multiple clones of phylogroup v is negligible, as the frequency of this phylogroup is much smaller than that of phylogroup k . We can deduce the proportion of empty

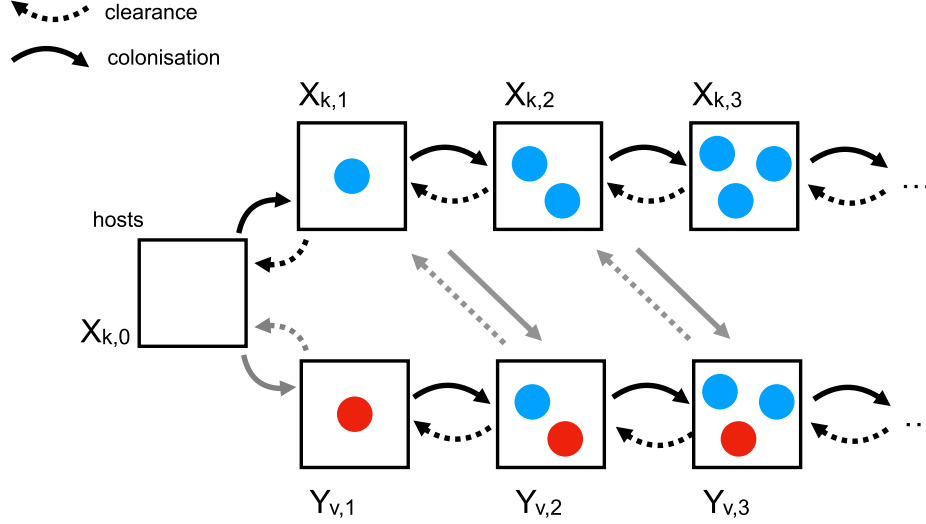


Figure S6: Potential transitions between host states (squares) because of clearance or colonization of either resident clones (blue dot) or invading clones (red dot). States with more than one invading clone are neglected. Arrows correspond to colonization (solid line) or clearance (dashed line) of clones of the resident (black line) or invading (grey line) phylogroup.

hosts from the condition that $\sum_{i=0}^{\infty} X_{k,i} + \sum_{i=1}^{\infty} Y_{v,i} \approx \sum_{i=0}^{\infty} X_{k,i} = 1$.

The equilibrium with phylogroup k alone corresponds to $X_{k,n} = \hat{X}_{k,n}$ and $Y_{v,n} = 0 \forall n \geq 1$. The dynamics of the system linearized at this equilibrium are given by the matrix:

$$\begin{bmatrix} \mathbf{J}_k & \mathbf{S} \\ \mathbf{0} & \mathbf{J}_v \end{bmatrix}.$$

The matrix \mathbf{J}_k contains the partial derivatives of $\dot{X}_{k,n}$ with respect to $X_{k,n}$, evaluated at the monomorphic equilibrium $\{\hat{\mathbf{X}}_k, \mathbf{0}\}$. $\hat{\mathbf{X}}_k$ corresponds to the vector of $\hat{X}_{k,n}$ values $\forall n \geq 1$ and $\mathbf{0}$ the vector of 0s, i.e. the equilibrium for $Y_{v,n}$. By definition of the stable equilibrium, the spectral bound of this matrix \mathbf{J}_k is smaller than 0 (Hurford et al., 2010). The matrix \mathbf{J}_v contains the partial derivatives of $\dot{Y}_{v,n}$ with respect to $Y_{v,n}$, evaluated at the monomorphic equilibrium $\{\hat{\mathbf{X}}_k, \mathbf{0}\}$. The matrix \mathbf{S} contains the partial derivatives of $X_{k,n}$ with respect to $Y_{v,n}$, evaluated at the monomorphic equilibrium $\{\hat{\mathbf{X}}_k, \mathbf{0}\}$. The lower-left matrix is $\mathbf{0}$ because the derivatives of $\dot{Y}_{v,n}$ with respect to $X_{k,n}$ evaluated at the invader-free equilibrium are all 0. The stability of the equilibrium is therefore governed by the spectral bound of \mathbf{J}_v , noted $s(\mathbf{J}_v)$.

Phylogroup v is at a neutral equilibrium, $s(\mathbf{J}_v) = 0$ if it has the same characteristics as k , i.e. if $R_v = R_k$ (Hurford et al., 2010). The two phylogroups differ only in their basic reproduction number R_k and R_v , as the competition coefficient b and d are assumed to be identical for both phylogroups. Thus, phylogroup v can invade the host population if $R_v > R_k$, as $s(\mathbf{J}_v) > 0$ in this case. Conversely, it cannot invade if $R_v < R_k$ as then $s(\mathbf{J}_v) < 0$.

6.4. Impact of changes in colonization or clearance rates on bacterial evolution

We now investigate how non-phylogroup-specific changes affect the competitive relationship between phylogroups. Let us consider two phylogroups x and y so that $R_x > R_y$. We first consider a ‘multiplicative’ change, so that the clearance or colonization rate of every phylogroup is multiplied by a strictly positive constant factor, e.g. by defining new clearance rates $\mu'_{k,n,n_k} = \eta \times \mu_{k,n,n_k}$, leading to $R'_k = R_k/\eta$. Because η is strictly positive ($\eta > 1$ and $\eta < 1$ respectively correspond to an increase and a decrease of

the clearance rate), $R'_x > R'_y$ for any value of η . Therefore, the hierarchy between phylogroups remains unaffected by such changes. Modifying colonization rate instead of clearance would lead to the same result, as multiplying the clearance rate by η has the same impact on the colonization-to-clearance ratio as multiplying the colonization rate by $1/\eta$.

Then, we consider an ‘additive’ change, so that a constant factor is added to the clearance or colonization rate of every phylogroup, e.g. by defining new clearance rates $\mu''_{k,n,n_k} = \mu_{k,n,n_k} + \theta$. This change is tantamount to multiplying the clearance rate by a phylogroup-specific factor θ_k^m :

$$\mu''_{k,n,n_k} = \theta_k^m \mu_{k,n,n_k}, \quad \theta_k^m = \frac{\mu_{k,n,n_k} + \theta}{\mu_{k,n,n_k}}, \quad (24)$$

leading to $R''_k = R_k/\theta_k^m$. Following Eq 24, this factor θ_k^m is negatively correlated with μ_{k,n,n_k} , meaning that this change disproportionately affects phylogroups with a slow turnover. Therefore, there is a factor θ^{eq} corresponding to the additional clearance rate canceling out the difference between R_x and R_y :

$$\theta^{eq} = \frac{R_x - R_y}{R_y e^{a_x} - R_x e^{a_y}}. \quad (25)$$

The phylogroup hierarchy is maintained, i.e. $R''_x > R''_y$ if $\theta < \theta^{eq}$, but reversed, i.e. $R''_x < R''_y$ if $\theta > \theta^{eq}$. The value of θ^{eq} is positive only if $a_x < a_y$, meaning that an additional source of clearance ($\theta > 0$) can only change the phylogroup hierarchy if the phylogroup with the higher initial colonization-to-extinction rate R_k also has the lower initial clearance rate μ_{k,n,n_k} .

7. Impact of an external source of *E. coli* clones

Our model assumes that any colonizing clone is directly transmitted from another host in the population. This implies two hypotheses: that transmission of *E. coli* is solely human-to-human, and that the host population is completely isolated. The first hypothesis ignores the potential role of the environment, which harbors various strains of *E. coli* (Money et al., 2010, Ludden et al., 2019) also susceptible to be transmitted to humans. The second hypothesis would be invalidated by rare transmissions of *E. coli* from hosts belonging to another population. Both hypotheses can be relaxed by modeling colonization as depending on a mixture of human-to-human within-population transmission and transmission from an external source also providing clones.

Following Levin et al. (1997), we considered a new colonization rate $\lambda_{k,n}^r(t)$, which depended in proportion $(1 - q)$ on transmission from other hosts in the population, and in proportion q on influx from this constant reservoir:

$$\lambda_{k,n}^r(t) = \frac{qF_k^r + (1 - q)F_k(t)}{e^{c_k + dn}},$$

with F_k^r the frequency of phylogroup k in the reservoir. This rate can also be expressed as the sum of two rates respectively corresponding to transmission from the reservoir and transmission within the population:

$$\lambda_{k,n}^r(t) = \underbrace{qF_k^r e^{-(c_k + dn)}}_{\text{reservoir}} + \underbrace{(1 - q)F_k(t) e^{-(c_k + dn)}}_{\text{within population}},$$

with the within-population part of the colonization rate corresponding to $(1 - q)\lambda_{k,n}(t)$. By definition, the contribution of the reservoir corresponds to a minimum colonization rate for phylogroup k , i.e. $\lambda_{k,n}^r(t) \geq qF_k^r e^{-(c_k + dn)}$. This threshold is always positive for any value of $q > 0$ and $F_k^r > 0$. Hence, completely eliminating a phylogroup becomes impossible because of the constant influx from the reservoir.

Using discrete-time simulations of the model with this colonization rate, we investigated how this external source impacted coexistence between two strains, by varying the proportion q of clones provided by the reservoir between 0 (no reservoir) to 1 (only the reservoir). We consider here the same case scenario as the one presented in the main text of the article (see the Results section): a couple of

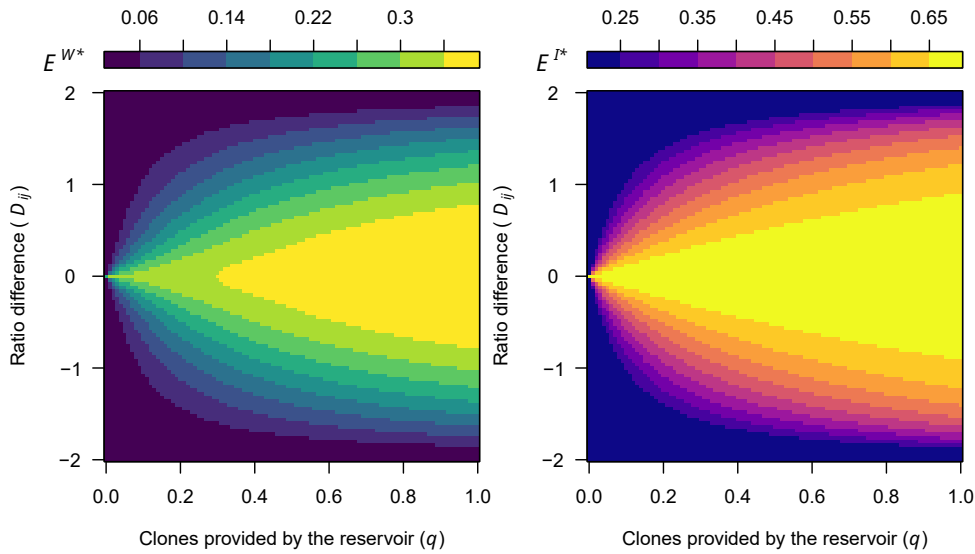


Figure S7: Asymptotic within host E^{W*} (A) and inter-hosts E^{I*} (B) coexistence as a function of the ratio difference $D_{i,j}$ and the proportion of clones provided by the reservoir q . For these simulations, $\bar{R}_{i,j} = 2$, $b = 0$, $d = 0$, $E^{W*} \in]0, 1]$ and $E^{I*} \in]0, \ln(2)]$.

phylogroups i and j characterized by an average colonization-to-clearance ratio $\overline{R_{i,j}} = (R_i + R_j) / 2$ and a ratio difference $D_{i,j} = R_i - R_j$. We considered both phylogroups present at the same frequency in the reservoir $F_i^r = F_j^r = 1/2$, and computed the asymptotic within-host coexistence E^{W*} and the asymptotic inter-host coexistence E^{I*} , with the same definition as presented in the main text of the article.

Results show that the reservoir facilitated coexistence (Fig S7). The range of values of $D_{i,j}$ values for which coexistence was stable widened dramatically as the proportion of colonization depending on the reservoir q increased. This was the case both for the within-host coexistence E^{W*} and the inter-host coexistence E^{I*} . These results show that even a reservoir with a somewhat limited impact on colonization would enable the coexistence of phylogroups with substantial differences in their colonization-to-clearance ratios. However, this effect would be mitigated if the fitter phylogroup were also the most abundant in the reservoir, e.g. if $R_i > R_j$ and $F_i^r > F_j^r$.

8. Supplementary figures

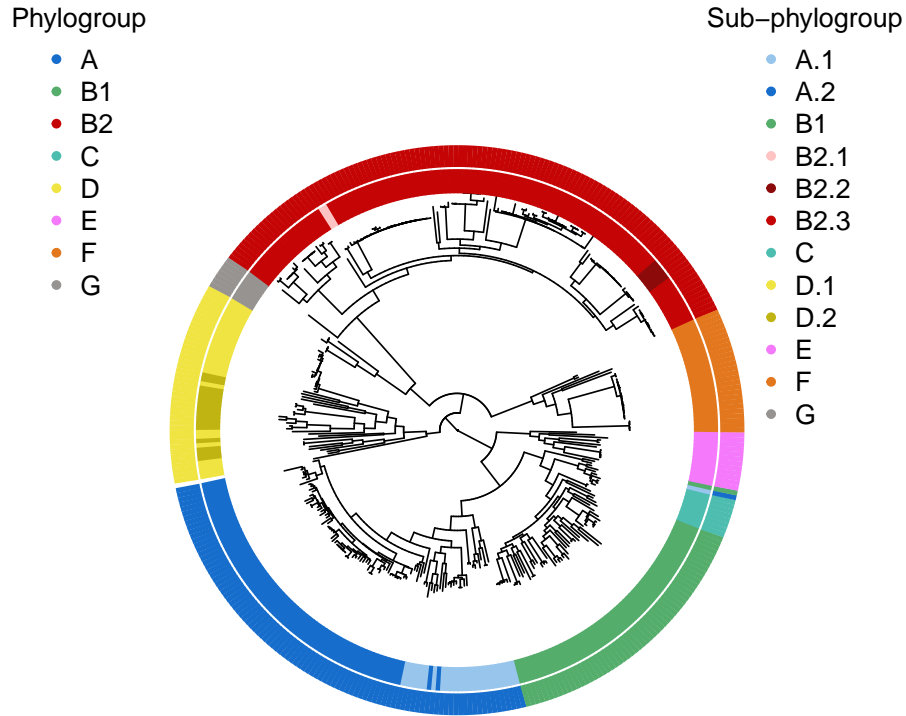


Figure S8: Phylogenetic tree of the commensal bacterial sequences of the additional dataset, with coloured rings indicating the phylogroup (outer) and subgroup (inner). The phylogenetic tree of the core genome was inferred with IQ-TREE multicore version 2.4.0 (Minh et al., 2020), with substitution model GTR+F+I+G (general time reversible with empirical frequencies, a proportion of invariable site, and a gamma distribution of rates). We rooted the tree using an outgroup of three clade sequences (two from clade IV and one from clade V). The phylogroups and sub-phylogroups were determined by Clermont PCR (Clermont et al., 2013) or in silico Clermont typing depending on the sequences (Beghain et al., 2018). The two A.2 sequences in the middle of A.1 correspond to an event of acquisition of the *YjaA* gene in these sequences. The two sequences A.1 and B1 inserted at the base of the B1/C clade probably form another (new) group.

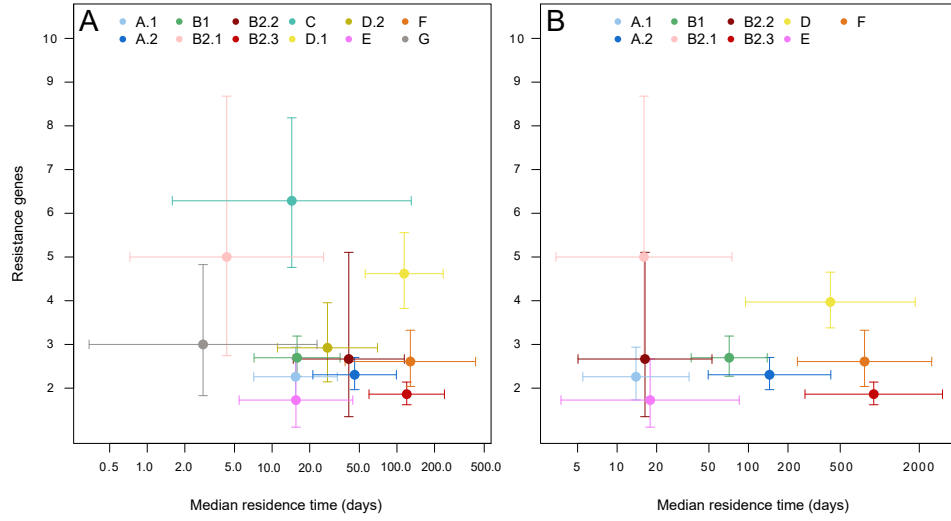


Figure S9: Antibiotic resistance genes as a function of median residence time for each group appearing in the French (A) and the USA (B) datasets, for the average host and average values of the quantitative explanatory factors in their respective datasets. Estimates (dots) are represented with their 95% confidence interval (bars)

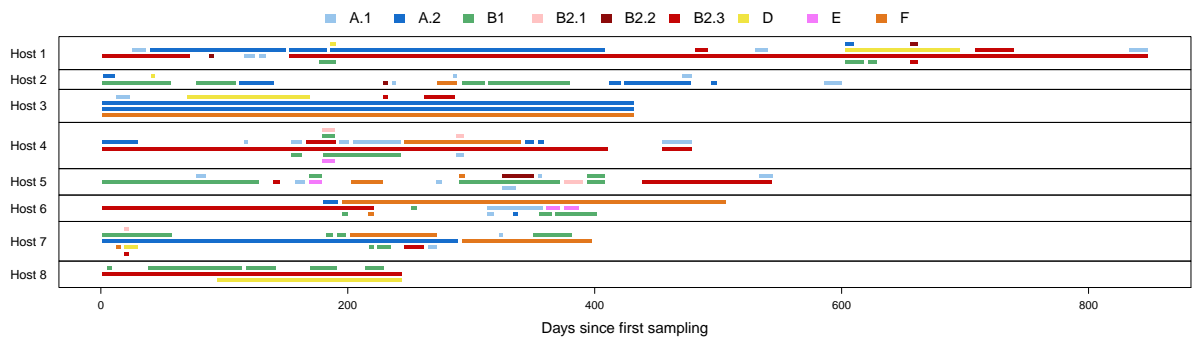


Figure S10: Presence of *E. coli* clones over time in each of the 8 hosts in the USA dataset. Each bar represents a single clone as defined in the study, starting either at 0, or midway between the first sampling including the clone and the previous one, and ending either at the date of the last sampling, or midway between the last sampling including the clone and the next one. Non-overlapping presence of several clones is represented on the same line for the sake of compactness, with no specific link between them.

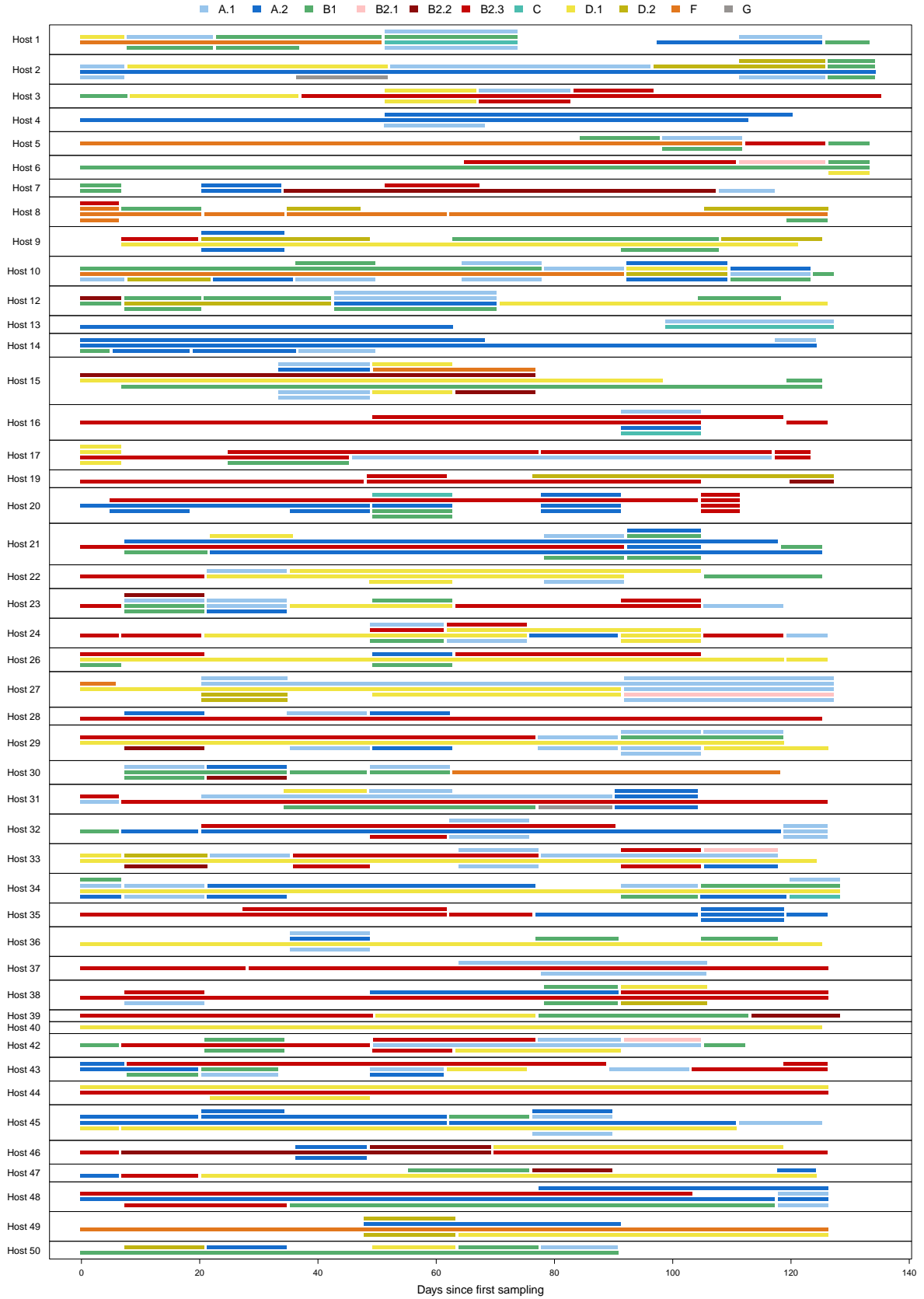


Figure S11: Presence of *E. coli* clones over time in each of the 46 hosts in the French dataset. Each bar represents a single clone as defined in the study, starting either at 0, or midway between the first sample including the clone and the previous one, and ending either at the date of the last sample, or midway between the last sample including the clone and the next one. Non-overlapping presence of several clones is represented on the same line for the sake of compactness, with no specific link between them.

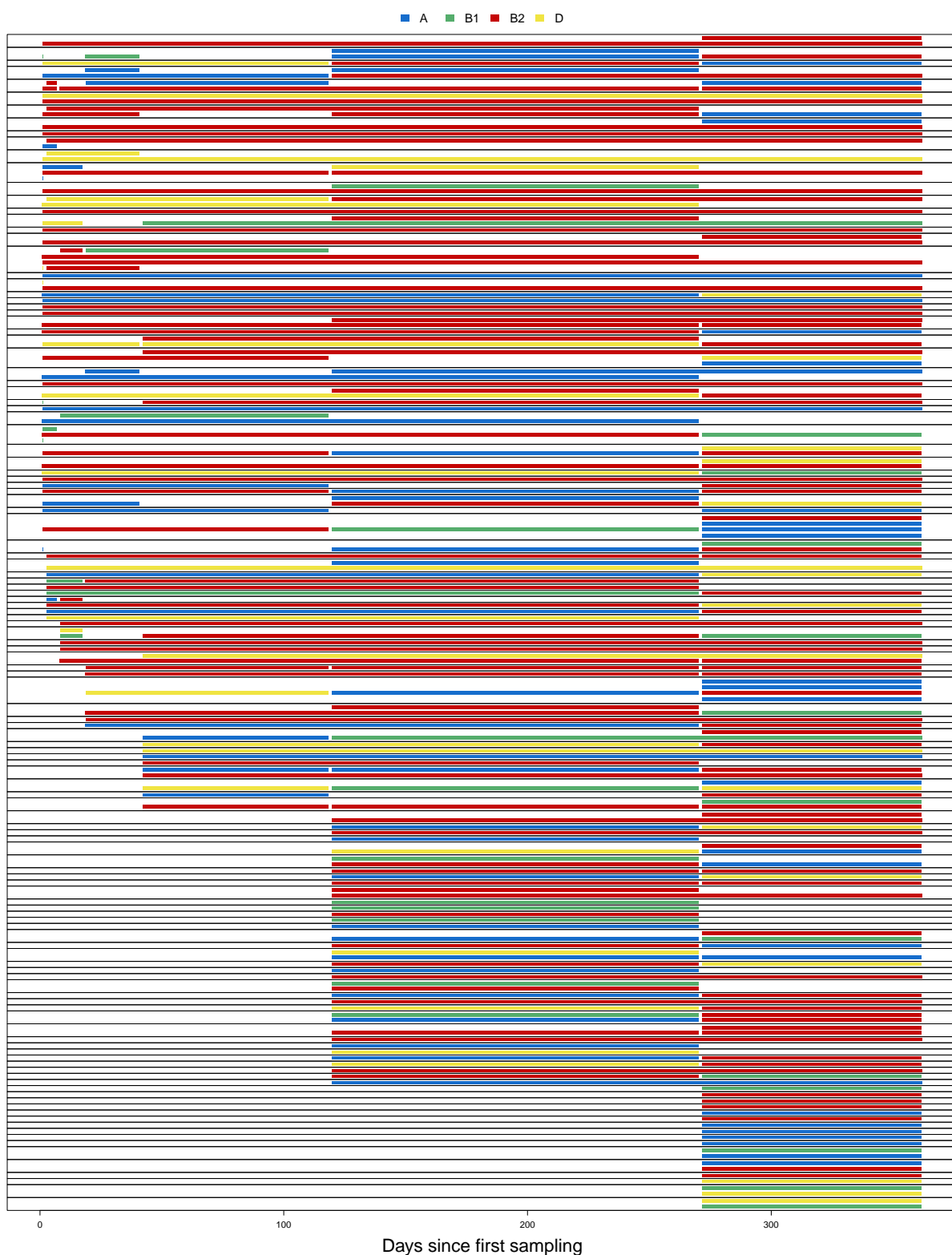


Figure S12: Presence of *E. coli* clones over time in each host of the infant (Sweden) dataset. Each bar represents a single clone as defined in the study, starting either at 0, or midway between the first sample including the clone and the previous one, and ending either at the date of the last sample, or midway between the last sample including the clone and the next one. Non-overlapping presence of several clones is represented on the same line for compactness, with no specific link between them. Hosts are separated by black lines.

9. Supplementary tables

9.1. AFT models for residence times

9.1.1 French dataset

Model	Distribution	AIC	Δ AIC
$Y \sim hst + grp + ncl + ngrp$	log-logistic	622.501	0.000
$Y \sim hst + grp + ncl + ngrp + \log(dns)$	log-logistic	623.848	1.346
$Y \sim hst + grp + ncl$	log-logistic	627.530	5.028
$Y \sim hst + grp + ncl + \log(dns)$	log-logistic	628.956	6.454

Table S1: AIC and Δ AIC of AFT models of residence for the French dataset, with Δ AIC<10, with either an exponential, Weibull or log-logistic distribution. The factors included are noted as follows: *hst* (host), *grp* (phylogroup), *ncl* (number of clones), *ngrp* (number of clones of the same phylogroup) and *dns* (total density).

Coefficient	Z score	P value
intercept (host 03/group D.2)	7.578	< 0.001
host 01	0.073	0.942
host 02	1.427	0.154
host 04	1.754	0.080
host 05	-1.960	0.050
host 06	-0.182	0.856
host 07	0.273	0.785
host 08	-1.619	0.105
host 09	1.362	0.173
host 10	-1.802	0.072
host 12	0.290	0.772
host 14	-1.059	0.290
host 15	-0.185	0.853
host 16	0.348	0.728
host 17	0.822	0.411
host 19	1.008	0.313
host 20	0.386	0.699
host 21	1.546	0.122
host 22	0.714	0.475
host 23	-0.583	0.560
host 24	-1.171	0.242
host 26	-0.149	0.881
host 27	0.040	0.968
host 28	-1.774	0.076
host 29	-0.200	0.841
host 30	-1.151	0.250
host 31	2.556	0.011
host 32	1.712	0.087
host 33	-1.125	0.261
host 34	0.322	0.748
host 35	-1.000	0.317
host 36	-1.015	0.310
host 37	2.701	0.007
host 38	-1.020	0.308
host 39	-1.041	0.298
host 42	-0.553	0.580
host 43	-0.561	0.575
host 44	0.357	0.721

Coefficient	<i>Z</i> score	<i>P</i> value
host 45	1.744	0.081
host 46	0.806	0.420
host 47	0.616	0.538
host 48	2.868	0.004
host 49	0.217	0.828
host 50	-2.217	0.027
group A.1	-2.311	0.021
group A.2	0.825	0.409
group B1	-1.886	0.059
group B2.1	-1.981	0.048
group B2.2	0.496	0.620
group B2.3	2.893	0.004
group C	-0.930	0.352
group D.	3.303	0.001
group E	-1.437	0.151
group F	3.230	0.001
group G	-3.336	0.001
number of clones	-4.218	< 0.001
number of clones of the group	-2.654	0.008
Log. of scale	-10.912	< 0.001

Table S2: Coefficients of the selected AFT model of residence for the French dataset, with their associated *Z* scores and *P* values. The model intercept uses host 03 and phylogroup D.2 as reference, as they are the closest to the average estimate among hosts and among groups, respectively. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.1.2 USA dataset

Model	Distribution	AIC	Δ AIC
$Y \sim hst + grp + ncl + ngrp + log(dscl)$	exponential	469.141	0.000
$Y \sim hst + grp + ncl + ngrp + log(dns) + log(dscl)$	exponential	470.128	0.987
$Y \sim hst + grp + ncl + ngrp + log(dscl)$	Weibull	470.555	1.414
$Y \sim hst + grp + ncl + ngrp$	exponential	471.031	1.891
$Y \sim hst + grp + ncl + log(dscl)$	exponential	471.392	2.252
$Y \sim hst + grp + ncl + ngrp + log(dns) + log(dscl)$	Weibull	471.402	2.261
$Y \sim hst + grp + ncl + ngrp + log(dns)$	exponential	471.890	2.749
$Y \sim hst + grp + ncl + log(dscl)$	Weibull	472.579	3.438
$Y \sim hst + grp + ncl + ngrp$	Weibull	472.610	3.469
$Y \sim hst + grp + ncl + log(dns) + log(dscl)$	exponential	472.991	3.850
$Y \sim hst + grp + ncl + ngrp + log(dns) + log(dscl)$	log-logistic	472.998	3.857
$Y \sim hst + grp + ncl$	exponential	473.381	4.240
$Y \sim hst + grp + ncl + ngrp + log(dns)$	Weibull	473.468	4.327
$Y \sim hst + grp + ncl + log(dns)$	exponential	473.742	4.601
$Y \sim hst + grp + ncl + log(dns) + log(dscl)$	Weibull	474.047	4.906
$Y \sim hst + grp + ncl$	Weibull	474.763	5.622
$Y \sim hst + grp + ncl + log(dns)$	Weibull	475.144	6.003
$Y \sim hst + grp + ngrp + log(dns) + log(dscl)$	exponential	475.178	6.037
$Y \sim hst + grp + ncl + log(dns) + log(dscl)$	log-logistic	475.351	6.210
$Y \sim hst + grp + ncl + ngrp + log(dscl)$	log-logistic	475.829	6.688
$Y \sim hst + grp + ngrp + log(dns) + log(dscl)$	Weibull	475.990	6.849
$Y \sim hst + grp + ncl + log(dscl)$	log-logistic	477.349	8.209

Table S3: AIC and Δ AIC of AFT models of residence for the USA dataset, with Δ AIC<10, with either an exponential, Weibull or log-logistic distribution. The factors included are noted as follows: *hst* (host), *grp* (phylogroup), *ncl* (number of clones), *ngrp* (number of clones of the same phylogroup), *dns* (total density) and *dscl* (total density of the clone).

Coefficient	<i>Z</i> score	<i>P</i> value
intercept (host 5/group A.2)	8.661	< 0.001
host 1	3.007	0.003
host 2	-0.926	0.355
host 3	-0.141	0.888
host 4	-1.985	0.047
host 6	1.311	0.190
host 7	-3.822	< 0.001
host 8	1.810	0.070
group A.1	-3.496	< 0.001
group B1	-0.990	0.322
group B2.1	-1.136	0.256
group B2.2	-3.103	0.002
group B2.3	4.916	< 0.001
group D	1.618	0.106
group E	-2.044	0.041
group F	3.679	< 0.001
number of clones	-3.444	< 0.001
number of clones of the group	-2.135	0.033
log. of clone density	1.983	0.047

Table S4: Coefficients of the selected AFT model of residence for the USA dataset, with their associated *Z* scores and *P* values. The model intercept uses host 5 and phylogroup A.2 as reference, as they are the closest to the average estimate among hosts and among groups, respectively. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.1.3 Infant dataset

Model	Distribution	AIC	Δ AIC
$Y \sim grp + ncl$	Weibull	303.041	0.000
$Y \sim grp + ncl + \log(dns)$	Weibull	304.851	1.810
$Y \sim grp + ncl + \log(dscl)$	Weibull	304.907	1.865
$Y \sim grp + ncl + ngrp$	Weibull	304.976	1.935
$Y \sim grp + ngrp$	Weibull	305.398	2.357
$Y \sim grp$	Weibull	305.743	2.701
$Y \sim grp + ncl + ngrp + \log(dns)$	Weibull	306.804	3.763
$Y \sim grp + ncl + \log(dns) + \log(dscl)$	Weibull	306.843	3.802
$Y \sim grp + ncl + ngrp + \log(dscl)$	Weibull	306.860	3.818
$Y \sim grp + ngrp + \log(dns)$	Weibull	307.323	4.282
$Y \sim grp + ngrp + \log(dscl)$	Weibull	307.366	4.324
$Y \sim grp + \log(dns) + \log(dscl)$	Weibull	307.556	4.515
$Y \sim grp + \log(dscl)$	Weibull	307.607	4.566
$Y \sim grp + \log(dns)$	Weibull	307.639	4.598
$Y \sim grp + ngrp + \log(dns) + \log(dscl)$	Weibull	308.373	5.331
$Y \sim grp + ncl + ngrp + \log(dns) + \log(dscl)$	Weibull	308.793	5.752

Table S5: AIC and Δ AIC of AFT models of residence for the infant dataset, with Δ AIC<10, with either an exponential, Weibull or log-logistic distribution. The factors included are noted as follows: *hst* (host), *grp* (phylogroup), *ncl* (number of clones), *ngrp* (number of clones of the same phylogroup), *dns* (total density) and *dscl* (total density of the clone).

9.2. AFT models for colonization rates

9.2.1 French dataset

Model	Distribution	AIC	Δ AIC
$Y \sim hst + grp + ncl$	Weibull	436.443	0.000
$Y \sim hst + grp + ncl + ngrp$	Weibull	437.722	1.279
$Y \sim hst + grp + ncl + \log(dns)$	Weibull	437.987	1.544
$Y \sim hst + grp + ncl + ngrp + \log(dns)$	Weibull	439.252	2.809
$Y \sim hst + grp + ncl$	log-logistic	440.191	3.748
$Y \sim hst + grp + ncl + ngrp$	log-logistic	441.242	4.798
$Y \sim hst + grp + ncl + \log(dns)$	log-logistic	441.395	4.952
$Y \sim hst + grp + ncl + ngrp + \log(dns)$	log-logistic	442.435	5.992

Table S6: AIC and Δ AIC of AFT models of colonization for the French dataset, with Δ AIC<10, with either an exponential, Weibull or log-logistic distribution. The factors included are noted as follows: *hst* (host), *grp* (phylogroup), *ncl* (number of clones), *ngrp* (number of clones of the same phylogroup) and *dns* (total density).

Coefficient	Z score	P value
intercept (host 21/group D.2)	9.157	< 0.001
host 01	-0.010	0.992
host 02	0.689	0.491
host 03	-0.915	0.360
host 05	-1.558	0.120
host 06	-0.761	0.447
host 07	0.548	0.583
host 08	0.713	0.476
host 09	-0.099	0.922
host 10	-1.584	0.113
host 12	-0.535	0.592
host 14	1.302	0.193
host 15	0.367	0.714
host 16	1.027	0.304
host 17	1.945	0.052
host 19	0.782	0.434
host 20	-0.487	0.626
host 22	-0.166	0.868
host 23	-1.212	0.226
host 24	-1.029	0.303
host 26	2.978	0.003
host 27	0.265	0.791
host 28	-0.803	0.422
host 29	-0.895	0.371
host 30	-2.247	0.025
host 31	-1.388	0.165
host 32	1.183	0.237
host 33	-1.704	0.088
host 34	1.903	0.057
host 35	-0.056	0.956
host 36	0.162	0.871
host 37	-0.828	0.408
host 38	0.275	0.784
host 39	-0.107	0.914
host 42	-1.458	0.145
host 43	-0.981	0.326
host 45	2.884	0.004
host 46	0.019	0.985

Coefficient	<i>Z</i> score	<i>P</i> value
host 47	-0.087	0.931
host 48	2.605	0.010
host 49	0.308	0.758
host 50	-0.698	0.485
group A.1	1.309	0.190
group A.2	6.781	< 0.001
group B1	5.493	< 0.001
group B2.1	-4.931	< 0.001
group B2.2	-2.983	0.003
group B2.3	9.335	< 0.001
group C	-3.030	0.002
group D.1	1.492	0.136
group E	-2.771	0.006
group F	1.240	0.215
group G	-2.371	0.018
number of clones	-9.493	< 0.001
Log. of scale	-12.918	< 0.001

Table S7: Coefficients of the selected AFT model of colonization for the French dataset, with their associated *Z* scores and *P* values. The model intercept uses host 21 and phylogroup D.2 as reference, as they are the closest to the average estimate among hosts and among groups, respectively. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.2.2 USA dataset

Model	Distribution	AIC	Δ AIC
$Y \sim hst + grp + ncl$	Weibull	481.535	0.000
$Y \sim hst + grp + ncl + ngrp$	Weibull	481.720	0.185
$Y \sim hst + grp + ncl + \log(dns)$	Weibull	482.886	1.350
$Y \sim hst + grp + ncl + ngrp + \log(dns)$	Weibull	483.002	1.466
$Y \sim hst + grp + ncl + ngrp$	log-logistic	489.479	7.944
$Y \sim hst + grp + ncl$	log-logistic	489.599	8.064
$Y \sim hst + grp + ncl + ngrp + \log(dns)$	log-logistic	491.254	9.719
$Y \sim hst + grp + ncl + \log(dns)$	log-logistic	491.473	9.938

Table S8: AIC and Δ AIC of AFT models of colonization for the USA dataset, with Δ AIC<10, with either an exponential, Weibull or log-logistic distribution. The factors included are noted as follows: *hst* (host), *grp* (phylogroup), *ncl* (number of clones), *ngrp* (number of clones of the same phylogroup), *dns* (total density) and *dscl* (total density of the clone).

Coefficient	<i>Z</i> score	<i>P</i> value
intercept (host 6/group B1)	17.879	< 0.001
host 1	1.698	0.090
host 2	-1.475	0.140
host 3	2.950	0.003
host 4	-1.451	0.147
host 5	-2.826	0.005
host 7	2.445	0.015
host 8	2.612	0.009
group A.1	-6.219	< 0.001
group A.2	3.874	< 0.001
group B2.1	-10.459	< 0.001
group B2.2	-7.768	< 0.001
group B2.3	8.472	< 0.001
group D	1.686	0.092
group E	-6.837	< 0.001
group F	-0.824	0.405
number of clones	-9.775	< 0.001
Log. of scale	-4.987	< 0.001

Table S9: Coefficients of the selected AFT model of colonization for the USA dataset, with their associated *Z* scores and *P* values. The model intercept uses host 6 and phylogroup B1 as reference, as they are the closest to the average estimate among hosts and among groups, respectively. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.2.3 Infant dataset

Model	Distribution	AIC	Δ AIC
$Y \sim grp + ncl$	Weibull	205.132	0.000
$Y \sim grp + ncl + log(dns)$	Weibull	206.868	1.737
$Y \sim grp + ncl + ngrp$	Weibull	207.047	1.915
$Y \sim grp$	Weibull	207.242	2.110
$Y \sim grp + ncl + ngrp + log(dns)$	Weibull	208.687	3.556
$Y \sim grp + log(dns)$	Weibull	208.805	3.674
$Y \sim grp + ngrp$	Weibull	208.942	3.810
$Y \sim grp + ngrp + log(dns)$	Weibull	210.304	5.172

Table S10: AIC and Δ AIC of AFT models of colonization for the USA dataset, with Δ AIC<10, with either an exponential, Weibull or log-logistic distribution. The factors included are noted as follows: *hst* (host), *grp* (phylogroup), *ncl* (number of clones), *ngrp* (number of clones of the same phylogroup) and *dns* (total density).

9.3. *Cox models for residence times*

9.3.1 French dataset

Coefficient	<i>Z</i> score	<i>P</i> value
group A.1	0.326	0.744
group A.2	-0.609	0.542
group B1	0.904	0.366
group B2.1	1.587	0.113
group B2.2	-0.543	0.587
group B2.3	-2.685	0.007
group C	0.138	0.891
group D.1	-3.108	0.002
group E	0.967	0.333
group F	-1.761	0.078
group G	1.859	0.063
number of clones	2.463	0.014
number of clones of the group	3.431	0.001

Table S11: Coefficients of the Cox model of residence for the French dataset, with their associated *Z* scores and *P* values. The model is stratified by host, as this factor did not respect the proportional hazard hypothesis. The model intercept uses phylogroup D.2 as the corresponding AFT model. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.3.2 USA dataset

Coefficient	<i>Z</i> score	<i>P</i> value
group A.1	2.214	0.027
group B1	0.192	0.848
group B2.1	1.045	0.296
group B2.2	1.648	0.099
group B2.3	-2.535	0.011
group D	-1.624	0.104
group E	0.955	0.340
group F	-3.093	0.002
number of clones	3.457	0.001
number of clones of the group	2.068	0.029
log. of clone density	-2.271	0.023

Table S12: Coefficients of the Cox model of residence for the USA dataset, with their associated *Z* scores and *P* values. The model is stratified by host, as this factor did not respect the proportional hazard hypothesis. The model intercept uses phylogroup A.2 as the corresponding AFT model. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.4. *Cox models for colonization rates*

9.4.1 French dataset

Coefficient	<i>Z</i> score	<i>P</i> value
group A.1	-0.320	0.749
group A.2	-4.056	< 0.001
group B1	-2.421	0.016
group B2.1	0.002	0.999
group B2.2	0.005	0.996
group B2.3	-5.708	< 0.001
group C	0.001	0.999
group D.1	-0.540	0.589
group E	2.619	0.009
group F	-2.655	0.008
group G	0.001	0.999
number of clones	5.403	< 0.001

Table S13: Coefficients of Cox model of colonization for the French dataset, with their associated *Z* scores and *P* values. The model is stratified by host, as this factor did not respect the proportional hazard hypothesis. The model intercept uses phylogroup D.2 as the corresponding AFT model. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

9.4.2 USA dataset

Coefficient	<i>Z</i> score	<i>P</i> value
group A.1	3.511	< 0.001
group B1	-2.338	0.010
group B2.1	6.794	< 0.001
group B2.2	3.476	< 0.001
group B2.3	-3.498	< 0.001
group D	-1.211	0.113
group E	6.687	< 0.001
group F	-0.811	0.209
number of clones	3.326	< 0.001

Table S14: Coefficients of the Cox model of colonization for the USA dataset, with their associated *Z* scores and *P* values. The model is stratified by host, as this factor did not respect the proportional hazard hypothesis. The model intercept uses phylogroup A.2 as the corresponding AFT model. The total number of clones and the number of clones of the same phylogroup were scaled prior to running the model, by subtracting their mean value and dividing them by their standard deviation.

References

- J. Beghain, A. Bridier-Nahmias, H. Le Nagard, E. Denamur, and O. Clermont. ClermonTyping: An easy-to-use and accurate in silico method for Escherichia genus strain phylotyping. *Microb Genom*, 4(7):e000192, 2018. ISSN 2057-5858. doi: 10.1099/mgen.0.000192.
- J. Burgaya, J. Marin, G. Royer, B. Condamine, B. Gachet, O. Clermont, F. Jaureguy, C. Burdet, A. Lefort, and V. 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.
- F. Caméléna, A. Birgy, Y. Smail, C. Courroux, P. Mariani-Kurkdjian, S. Le Hello, S. Bonacorsi, and P. 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.
- O. Clermont, S. Bonacorsi, and E. 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.
- O. Clermont, J. K. Christenson, E. Denamur, and D. M. Gordon. The clermont Escherichia coli phylotyping method revisited: Improvement of specificity and detection of new phylo-groups. *Environ Microbiol Rep*, 5(1):58–65, 2013. doi: 10.1111/1758-2229.12019.
- O. Clermont, O. V. Dixit, B. Vangchhia, B. Condamine, S. Dion, A. Bridier-Nahmias, E. Denamur, and D. Gordon. Characterization and rapid identification of phylogroup G in Escherichia coli, a lineage with high virulence and antibiotic resistance potential. *Environ Microbiol*, 21(8):3107–3117, 2019. ISSN 1462-2912. doi: 10.1111/1462-2920.14713.
- O. Clermont, B. Condamine, S. Dion, D. M. Gordon, and E. Denamur. The e phylogroup of Escherichia coli is highly diverse and mimics the whole E. coli species population structure. *Environ Microbiol*, 23(11):7139–7151, 2021. doi: 10.1111/1462-2920.15742.
- P. Escobar-Páramo, K. Grenet, A. Le Menac’h, L. Rode, E. Salgado, C. Amorin, S. Gouriou, B. Picard, M. C. Rahimy, and A. Andreumont. Large-scale population structure of human commensal Escherichia coli isolates. *Appl Environ Microbiol*, 70(9):5698–5700, 2004. doi: 10.1128/AEM.70.9.5698-5700.2004.
- E. L. Hannah, J. R. Johnson, F. Angulo, B. Haddadin, J. Williamson, and M. H. Samore. Molecular analysis of antimicrobial-susceptible and-resistant Escherichia coli from retail meats and human stool and clinical specimens in a rural community setting. *Foodborne Pathog Dis*, 6(3):285–295, 2009. doi: 10.1089/fpd.2008.0176.
- A. Hurford, D. Cownden, and T. Day. Next-generation tools for evolutionary invasion analyses. *J R Soc Interface*, 7(45):561–571, 2010. ISSN 1742-5689. doi: 10.1098/rsif.2009.0448.
- J. R. Johnson, K. Owens, A. Gajewski, and M. A. Kuskowski. Bacterial characteristics in relation to clinical source of Escherichia coli isolates from women with acute cystitis or pyelonephritis and uninfected women. *J Clin Microbiol*, 43(12):6064–6072, 2005. doi: 10.1128/jcm.43.12.6064-6072.2005.
- J. R. Johnson, M. R. Sannes, C. Croy, B. Johnston, C. Clabots, M. A. Kuskowski, J. Bender, K. E. Smith, P. L. Winokur, and E. A. Belongia. Antimicrobial drug-resistant Escherichia coli from humans and poultry products, Minnesota and Wisconsin, 2002–2004. *Emerg Infect Dis*, 13(6):838, 2007. doi: 10.3201/eid1306.061576.

- N. Karami, A. Martner, V. I. Enne, S. Swerkersson, I. Adlerberth, and A. E. Wold. Transfer of an ampicillin resistance gene between two *Escherichia coli* strains in the bowel microbiota of an infant treated with antibiotics. *J Antimicrob Chemother*, 60(5):1142–1145, 2007. doi: 10.1093/jac/dkm327.
- B. Levin, M. Lipsitch, V. Perrot, S. Schrag, R. Antia, L. Simonsen, N. Moore Walker, and F. Stewart. The population genetics of antibiotic resistance. *Clin Infect Dis*, 24(Supplement_1):S9–S16, 1997. doi: 10.1093/clinids/24.supplement_1.s9.
- C. M. Logue, C. Doetkott, P. Mangiamale, Y. M. Wannemuehler, T. J. Johnson, K. A. Tivendale, G. Li, J. S. Sherwood, and L. K. Nolan. Genotypic and phenotypic traits that distinguish neonatal meningitis-associated *Escherichia coli* from fecal *E. coli* isolates of healthy human hosts. *Appl Environ Microbiol*, 78(16):5824–5830, 2012. ISSN 0099-2240. doi: 10.1128/AEM.07869-11.
- C. Ludden, K. E. Raven, D. Jamroz, T. Gouliouris, B. Blane, F. Coll, M. de Goffau, P. Naydenova, C. Horner, and J. 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.
- J. N. Martinson, N. V. Pinkham, G. W. Peters, H. Cho, J. Heng, M. Rauch, S. C. Broadaway, and S. 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.
- B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*, 37(5):1530–1534, 2020. doi: 10.1093/molbev/msaa015.
- B. 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.
- P. Money, A. Kelly, S. Gould, J. Denholm-Price, E. Threlfall, and M. 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.
- F. L. Nowrouzian, A. E. Wold, and I. 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.
- A. Ostblom, I. Adlerberth, A. E. Wold, and F. 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.
- M. R. Sannes, M. A. Kuskowski, K. Owens, A. Gajewski, and J. R. Johnson. Virulence factor profiles and phylogenetic background of *Escherichia coli* isolates from veterans with bacteremia and uninfected control subjects. *J Infect Dis*, 190(12):2121–2128, 2004. ISSN 1537-6613. doi: 10.1086/425984.
- D. Skurnik, D. Bonnet, C. Bernède-Bauduin, R. Michel, C. Guette, J.-M. Becker, C. Balaire, F. Chau, J. Mohler, and V. Jarlier. Characteristics of human intestinal *Escherichia coli* with changing environments. *Environ Microbiol*, 10(8):2132–2137, 2008. doi: 10.1111/j.1462-2920.2008.01636.x.
- L. Zhang, B. Foxman, and C. Marrs. Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group B2. *J Clin Microbiol*, 40(11):3951–3955, 2002. ISSN 0095-1137. doi: 10.1128/jcm.40.11.3951-3955.2002.