Abstract

Since its discovery, mobile DNA has fascinated researchers. In particular, many researchers have debated why insertion sequences persist in genomes and populations. While some authors think that insertion sequences persist only because of occasional beneficial effects on their hosts, others argue that horizontal gene transfer is strong enough to overcome the generally detrimental effects on their hosts. In this study, we model the long-term fate of a prokaryote cell population of which a small proportion of cells has been infected with one insertion sequence. Based on our model and the count distribution of insertion sequence IS\textsubscript{5} found in 525 fully sequenced proteobacterial genomes, we show that the fitness cost of IS\textsubscript{5} is so small that IS\textsubscript{5} is effectively neutral or only slightly detrimental. We also show that an IS\textsubscript{5} infection can persist and reach the empirically observed distribution if the rate of horizontal gene transfer is at least as large as the fitness cost, and that this rate is well within the rates of horizontal gene transfer observed in nature. In addition, we show that the time needed to reach the observed prevalence of IS\textsubscript{5} is unrealistically long for the fitness costs and horizontal gene transfer rates that we computed. Occasional beneficial effects may thus have played an important role in the fast spreading of IS\textsubscript{5}.

\textsuperscript{*}Institute of Evolutionary Biology and Environmental Studies, University of Zürich, Switzerland
\textsuperscript{†}Institute of Mathematics, University of Zürich, Switzerland
\textsuperscript{‡}Institute of Evolutionary Biology and Environmental Studies, University of Zürich, Switzerland
Estimating the Fitness Effect of the Insertion Sequence IS5

Introduction

Bacterial insertion sequences (ISs) are the simplest form of autonomous mobile DNA. They are short (800–2500 bp) DNA sequences typically consisting of one open reading frame that codes for the enzyme transposase which is needed for transposition. The open reading frame is flanked by short terminal inverted repeats which serve as recognition sites for the transposase. This enzyme usually excises the IS and inserts it elsewhere in the genome (conservative transposition), but occasionally it replicates the IS during this transposition process (replicative transposition) [Chandler and Mahillon, 2002]. An IS may get lost from a genome through excision. ISs have been assigned numbers, roughly in the order of their discovery: e.g. IS1A, IS5, IS630. Based on their internal structure and the inverted repeats, all ISs have been classified into 20 different families [Chandler and Mahillon, 2002, Mahillon et al., 2009]. The focus of our study, IS5, belongs to a rather heterogeneous family of ISs that is widely distributed among bacteria and archaea.

IS5 and all other ISs are inherited through vertical transmission. But they can also be horizontally transmitted by horizontal gene transfer (HGT) between prokaryotes, i.e. by natural transformation, by transduction through phages, and by conjugation through plasmids. The reported rates of transposition, excision and HGT are typically very low. Table 1 provides an overview over these rates.

Due to their transposition activity and the deletions, insertions and inversions through homologous recombination that are possible if more than one IS is present in a genome [Galas and Chandler, 1989, Kleckner, 1989, Schneider and Lenski, 2004], ISs pose a potential threat to their hosts, although occasional beneficial effects have also been reported [Hall, 1999, Schneider and Lenski, 2004]. Besides acting on their own, two ISs can also build a composite transposon, which consists of two copies of an IS that flank intermediary genes and transpose synchronously, thereby mobilising the intermediary genes. In this way, ISs are involved in transferring genes that confer resistance to antibiotics [Berg, 1989, Kleckner, 1989], genes that encode toxins [So and McCarthy, 1980], or genes with new metabolic functions [Top and Springael, 2003]. On the one hand, ISs therefore help to spread antibiotic resistance among pathogens and pose a public health threat, but
on the other hand, ISs are also valuable tools used in genetic engineering.

While most authors agree that harboring ISs in the genome is in general detrimental to the cell, there is disagreement about whether ISs persist because they are occasionally beneficial to their hosts [Blot, 1994, Shapiro, 1999, Schneider and Lenski, 2004] or because HGT is sufficiently strong to overcome selection against ISs due to their detrimental effects on their hosts [Dawkins, 1976, Doolittle and Sapienza, 1980, Orgel and Crick, 1980, Charlesworth et al., 1994, Nuzhdin, 1999].

In an earlier study, we used a stochastic, branching process model to show that even purely detrimental ISs can invade a host cell population and persist, provided that the HGT rate is larger than the fitness cost caused by the IS [Bichsel et al., 2010]. Based on our model, we showed that most IS infections do not persist and die out very quickly, and that those infections that do persist take a very long time to reach noticeable population size thresholds. While the branching process model is well suited to model the initial phase of an IS infection, it does not allow for interactions between cells, and is not useful for modeling the long-term effects of an IS infection. In this study, we therefore use a deterministic model based on a system of ordinary differential equations to examine whether purely detrimental ISs can persist. We then determine how large a fitness cost of an IS and a HGT rate would be needed to create the IS count distribution of IS5 that can be observed in proteobacterial genomes.

Data, Model, and Methods

Data

We obtained the genome sequences of 1447 fully sequenced prokaryote genomes from 542 genera that were available at NCBI on September 1, 2011 [NCBI, 2011]. We also obtained the sequences of one representative IS from each of the 20 known IS families from the IS Finder database [Mahillon et al., 2009]. We then used IScan [Wagner et al., 2007] to search the 1447 genome sequences (only chromosomes, no plasmids) for the 20 IS sequences. For later analysis, we needed independent IS count observations. We were therefore interested in ISs which have infected genomes over a large range of genera. Of all 20 ISs that have been examined, only 3 occur in more than 10 different genera. And
Estimating the Fitness Effect of the Insertion Sequence IS5

of these 3 ISs, only IS5 occurs often enough in these genera so that a random sample of
one genome per genus contains on average more than 10 infected genomes. To get more
dependable results in our statistical analysis, we therefore focused on IS5. It turned out
that IS5 (as most of the other 20 ISs we examined) can be found mainly in genomes from
proteobacteria: only 4 of 58 infected genomes do not belong to proteobacteria. We thus
restricted our IS5 count analysis to proteobacteria. In our data set, this phylum consists
of 525 genomes in 180 genera, where we have added Shigella to the genus Escherichia
[Lan and Reeves, 2002].

Model Design

Figure 1 shows the design of our model.

We assume an uninfected prokaryote host cell population living at carrying capacity $K$,
where $K$ is a cell population density. The prokaryote cells live in a well-mixed bulk
environment, and their normalized population density is given by $Z_0 = D_0/K$, where
$D_0$ is their density. The change of $Z_0$ over time is governed by the logistic equation
$\dot{Z}_0 = r(1 - Z_0)Z_0$, with the base population growth rate $r$. We set $r = 1$, so that one time
unit corresponds to the doubling time during the early exponential growth phase, and we
take this as the generation time of a cell. At the begin of an IS infection, each cell of a
very small proportion of cells (e.g. $10^{-6}$) is infected with one IS in its genome. We then
model the spread of the IS infection through the host cells and compute the equilibrium
distribution of the IS count per prokaryotic cell genome. To do so, we use a system of
ordinary differential equations for the normalized cell densities $Z_k = D_k/K$, where $D_k$ is
the density of cells carrying $k$ ISs in their genome. To keep the computation numerically
tractable, we limit the maximal number of ISs per infected cell to $l = 60$. This is not a
strong limitation, because only few genomes harbor more than 60 ISs, as other authors
have reported [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007]. We show in
the results section that no genome in our data set contains more than 60 copies of IS5, the
focus of our interest. Besides the base population growth rate $r$, our model contains the
following rate parameters: the base fitness effect $s$ of one IS, the replicative transposition
rate $u$ of one IS, the excision rate $e$ per IS, and the HGT rate $h$.

We allow for a nonlinear impact of an increasing IS count per cell on the cell’s fitness, its infectiousness to other cells, and its total replicative transposition rate. To this end, we model the fitness effect, the HGT rate and the total replicative transposition rate of all ISs in a cell as a power function of the cell’s IS count, with exponents $\alpha$, $\beta$ and $\gamma$, respectively. We choose $\alpha, \beta, \gamma \in \{0, 1, 2\}$, where an exponent of 0 reflects independence of the rate from the cell’s total IS count, an exponent of 1 reflects linear dependence, and an exponent of 2 reflects quadratic dependence of the rate from the cell’s total IS count. This is equivalent to a diminishing (exponent 0), a constant (exponent 1) and an increasing (exponent 2) effect per IS of an increasing IS count on the rate. For simplicity, we let the total excision rate increase linearly with the cell’s IS count, i.e. we assume that ISs are excised independently of each other.

Our data suggest that the number of infected cells in a population stays low compared to the total number of cells (see figure 2). This has also been observed before [Wagner, 2006, Touchon and Rocha, 2007]. To simplify our model, we therefore assume that infected cells are surrounded by uninfected cells only, and that no HGT occurs between infected cells. Furthermore, we assume that during HGT only one IS gets copied from an infected to an uninfected cell. This is justified by the observation that during transformation and transduction typically only small DNA fragments are transferred from one cell to another, and that ISs on a plasmid transferred during conjugation must first be inserted into the chromosome [Madigan et al., 2009, p. 297ff].
Model Analysis

Based on our model design shown in figure 1, we describe the dynamics of an IS infection with the following system of ordinary differential equations, where $Z = \sum_{k=0}^{l} Z_k \geq 0$:

\[
\dot{Z}_0 = r(1 - Z)Z_0 - h Z_0 \sum_{k=1}^{l} k^\beta Z_k + e Z_1 \\
\dot{Z}_1 = [r(1 - Z) + s] Z_1 + h Z_0 \sum_{k=1}^{l} k^\beta Z_k - u Z_1 - e Z_1 + 2e Z_2 \\
\dot{Z}_2 = [r(1 - Z) + 2s] Z_2 + u Z_1 - 2\gamma u Z_2 - 2e Z_2 + 3e Z_3 \\
\vdots \\
\dot{Z}_j = [r(1 - Z) + j^s] Z_j + (j - 1)^\gamma u Z_{j-1} - j^\gamma u Z_j - jep_j + (j + 1)e Z_{j+1} \\
\vdots \\
\dot{Z}_l = [r(1 - Z) + l^s] Z_l + (l - 1)^\gamma u Z_{l-1} - le Z_l
\]

This system has two obvious equilibrium solutions: the first one is $Z_0 = Z_1 = \ldots = Z_l = 0$, i.e. population extinction, and the second one is $Z_0 = 1$ and $Z_1 = \ldots = Z_l = 0$, i.e. IS extinction. We are more interested in equilibria where not all $Z_k$ for $k \in \{1, \ldots, l\}$ vanish.

In that case $Z > 0$, and using the proportions $p_k = Z_k/Z$ and their derivatives with respect to time,

\[
\dot{p}_k = \frac{\dot{Z}_k}{Z} - \frac{Z_k \cdot \dot{Z}}{Z^2} = \frac{1}{Z} \left( \dot{Z}_k - p_k \sum_{j=0}^{l} \dot{Z}_j \right) = \frac{1}{Z} \dot{Z}_k - p_k \left( r(1 - Z) - s \sum_{j=1}^{l} j^\alpha p_j \right),
\]

we define a new system of ordinary differential equations for $p_k$ ($k \in \{0, \ldots, l\}$) and for $Z$:

\[
p_0 = -h p_0 Z \sum_{k=1}^{l} k^\beta p_k + ep_1 - s p_0 \sum_{k=1}^{l} k^\alpha p_k \\
p_1 = sp_1 + h p_0 Z \sum_{k=1}^{l} k^\beta p_k - up_1 - ep_1 + 2ep_2 - s p_1 \sum_{k=1}^{l} k^\alpha p_k \\
p_2 = 2\alpha sp_2 + up_1 - 2\gamma up_2 - 2ep_2 + 3ep_3 - s p_2 \sum_{k=1}^{l} k^\alpha p_k \\
\vdots \\
p_j = j^\alpha sp_j + (j - 1)^\gamma up_{j-1} - j^\gamma up_j - jep_j + (j + 1)ep_{j+1} - s p_j \sum_{k=1}^{l} k^\alpha p_k \\
\vdots \\
p_l = l^\alpha sp_l + (l - 1)^\gamma up_{l-1} - lep_l - s p_l \sum_{k=1}^{l} k^\alpha p_k
\]
and

\[ \dot{Z} = \sum_{k=0}^{l} \dot{Z}_k = r(1 - Z)Z + sZ \sum_{k=1}^{l} k^\alpha p_k = \left( r(1 - Z) + s \sum_{k=1}^{l} k^\alpha p_k \right) Z. \quad (3) \]

Besides setting \( r = 1 \), we set the replicative transposition rate \( u \) and the excision rate \( e \) to one of two fixed parameter sets that together cover a range of realistic rates (see table 1). In the main text, we use \((u,e) = (10^{-7}, 10^{-10})\), and in the appendix we use \((u,e) = (10^{-9}, 10^{-11})\). To solve the system (2, 3), we define

\[ p = (p_0, \ldots, p_l)^T, S_\alpha(p) = \sum_{k=1}^{l} k^\alpha p_k, S_\beta(p) = \sum_{k=1}^{l} k^\beta p_k, \]

and the HGT parameter \( H(p, Z) = h S_\beta(p) Z \). Observe that \( H \geq 0 \). The differential equations for \( p_0, \ldots, p_l \) in (2) can now be written in vector notation as

\[ \dot{p} = M(s, H(p, Z)) \cdot p - s S_\alpha(p) p \quad (4) \]

where

\[
M(s, H) = \begin{pmatrix}
-H & e \\
H & s - u - e & 2e \\
u & 2^\alpha s - 2^\gamma u - 2e & 3e \\
 & \ldots & \ldots & \ldots \\
(j - 1)^\gamma u & j^\alpha s - j^\gamma u - je & (j + 1)e & \ldots & \ldots \\
 & \ldots & \ldots & \ldots \\
(l - 1)^\gamma u & l^\alpha s - l^\gamma u - le \\
\end{pmatrix}
\]

We get again the IS extinction equilibrium for \( p = e_0 = (1, 0, \ldots, 0)^T \), because \( S_\alpha(e_0) = S_\beta(e_0) = H(e_0, Z) = 0 \) for any \( Z > 0 \), and therefore \( M(s, 0) \cdot e_0 = 0 \), so that \( M(s, 0) \cdot p - s S_\alpha(p) p = 0 \). For all other equilibrium solutions \((p, Z)\) of (2, 3) that may exist, \( H = H(p, Z) \) and \( \lambda = s S_\alpha(p) \) must fulfill \( M(s, H) \cdot p = \lambda p \). We are therefore looking for non-negative eigenvectors of the matrix \( M(s, H) \) for \( H > 0 \) (\( H = 0 \) is not interesting, because it implies \( Z_1 = \ldots = Z_l = 0 \)).

\( M(s, H) \) for \( H > 0 \) is a Metzler-Leontief matrix, i.e. \( (M)_{ij} \geq 0 \) for \( i \neq j \) [Seneta, 1981, p. 45]. In addition, \( M \) is irreducible. Therefore, for any choice of \( H > 0 \), there exists an
eigenvalue \( \tau \in \mathbb{R} \) such that \( \tau > \text{Re}(\mu) \) for all other eigenvalues \( \mu \) of \( \mathbf{M} \), and there exists

a unique (up to multiples), strictly positive eigenvector \( \mathbf{q} \) associated with \( \tau \). \( \mathbf{q} \) can be

normed so that \( \|\mathbf{q}\|_1 = 1 \). Furthermore,

1. if \( \mathbf{M}(s, H) \cdot \mathbf{p} = \eta \mathbf{p} \) for a specific eigenvector \( \mathbf{p} \) with \( \sum_{k=0}^{l} p_k = 1 \), then \( (1, \ldots, 1) \cdot \mathbf{M}(s, H) \cdot \mathbf{p} = s S_\alpha(\mathbf{p}) = \lambda \) for all proportion vectors \( \mathbf{p} \) (see the differential

   equations (2) for \( \mathbf{p} \)),

   and therefore \( \tau = \lambda = s S_\alpha(\mathbf{q}) \).

We now have

\[
\dot{\mathbf{q}} = \mathbf{M}(s, H(\mathbf{q}, Z)) \cdot \mathbf{q} - s S_\alpha(\mathbf{q}) \mathbf{q} = 0,
\]

and therefore, if we set \( Z = 1 + \frac{s}{\tau} S_\alpha(\mathbf{q}) \), so that \( \dot{Z}(\mathbf{q}) = [r(1 - Z) + s S_\alpha(\mathbf{q})] \cdot Z = 0 \), the

pair \( (\mathbf{q}, Z) \) is an equilibrium solution of the system (2, 3) for the proportions \( \mathbf{p} \) and the

total population size \( Z \).

Note that it is hard to compute an equilibrium solution based directly on \( h, \beta, \) and \( s \),
because one then has to solve the differential equation system (2, 3). But it is much easier
to algebraically compute an equilibrium solution of (2, 3) for a given pair \( (s, H) \) with

\( H > 0 \) and then to find values of \( h = h_\beta \) for any \( \beta \in \{0, 1, 2\} \). Here are the computational

steps that are required:

1. Compute the unique eigenvector \( \mathbf{q} \) with \( \|\mathbf{q}\|_1 = 1 \) that corresponds to the (real)
eigenvalue \( \tau \) with the largest real part of the matrix \( \mathbf{M}(s, H) \).

2. Set \( Z = 1 + \frac{s}{\tau} S_\alpha(\mathbf{q}) = 1 + \frac{s}{\tau} \sum_{k=1}^{l} k^\alpha q_k \).

3. Compute \( h_\beta = \frac{H}{S_\beta(\mathbf{q}) Z} = \frac{H}{\sum_{k=1}^{l} k^\beta q_k Z} \) for \( \beta \in \{0, 1, 2\} \).

To assess the local stability of the equilibrium distribution \((q_0, \ldots, q_l)\), we first calculate

\( Z_j = q_j Z \) for \( j \in \{0, \ldots, l\} \). We then compute the eigenvalues of the Jacobian matrix

\[ J = \left( \frac{\partial f_j(Z_0, \ldots, Z_l)}{\partial Z_j} \right)_{i,j \in \{0, \ldots, l\}} \]

at the specific values of \( (s, h) \) and \( (\alpha, \beta, \gamma) \) used to compute
the equilibrium. Here, \( f_j(Z_1, \ldots, Z_l) \) is the right-hand side of the differential equation for

\( Z_i \) in the system (1). The equilibrium is locally stable if the real parts of all eigenvalues
are negative.

The global stability of the equilibrium is much more difficult to establish. We confine
ourselves to check whether the equilibrium is reached, starting from an initial cell population at carrying capacity infected with a small proportion of cells harboring one IS in their genome. To do so, we numerically solve the system (1) with the values of \((s, h)\) and \((\alpha, \beta, \gamma)\) used to compute the equilibrium. We have chosen the values \(10^{-9}\), \(10^{-6}\), and \(10^{-3}\) as proportions of initially infected cells.

To find values for \((\alpha, \beta, \gamma)\) and \((s, h)\) that lead to the best approximation of an observed IS\(5\) count distribution by our theoretical IS count distribution, we use a maximum likelihood method and compute maximum likelihood estimates of \((\alpha, \beta, \gamma)\) and \((s, h)\). We start by defining the likelihood function \(L\). Given the observed IS\(5\) counts \((c_0, \ldots, c_l)\) and the predicted IS count distribution \((q_0, \ldots, q_l)\) based on the parameters \(\alpha, \gamma, s\) and \(H\), the likelihood function is given by

\[
L(\alpha, \gamma, s, H) = q_0^{c_0} \cdot \cdots \cdot q_l^{c_l},
\]

and its (natural) logarithm is

\[
\ln(L(\alpha, \gamma, s, H)) = c_0 \cdot \ln(q_0) + \ldots + c_l \cdot \ln(q_l).
\]

As we can only numerically compute the vector of proportions \(q\) based on the parameters \(\alpha, \gamma, s,\) and \(H\) and cannot derive \(q\) in analytical form, we use the Nelder-Mead method [Nelder and Mead, 1965] to find the maximum log-likelihood in the parameter space \((s, H)\) for all pairs \((\alpha, \gamma)\) \(\in\) \(\{0, 1, 2\}^2\). Having found the maximum likelihood estimates \(\hat{s}\) and \(\hat{H}\) for the combination \((\hat{\alpha}, \hat{\gamma})\) of \(\alpha\) and \(\gamma\) that maximises the likelihood function \(L\), we then obtain the maximum likelihood estimate \(\hat{h} = \hat{h}_\beta\) for all values of \(\beta\) \(\in\) \(\{0, 1, 2\}\) by following the three computational steps described above, replacing \(s\) by \(\hat{s}\) and \(H\) by \(\hat{H}\) throughout.

We then use the bootstrap method with 1000 resamplings of an IS\(5\) count distribution to show the association between the fitness effect \(s\) and the HGT rate \(h\) and to compute the corresponding 95%-confidence intervals [Efron and Tibshirani, 1994, p. 170].

For the numerical analysis, we use Mathematica 8.0.0 [Wolfram, 2003].
Results

The IS\textsubscript{5} count distribution in proteobacterial cells is L-shaped

Figure 2 shows the IS\textsubscript{5} count distribution based on 525 fully sequenced, proteobacterial genomes. In the figure, the mean distribution over 1000 random samples is shown. Each sample consists of 180 genomes, one randomly chosen genome per proteobacterial genus. Averaging over 1000 random samples provides us with an approximation of the real IS\textsubscript{5} count distribution over proteobacterial genera. Furthermore the 1000 random samples give an impression of the uncertainty about the real IS\textsubscript{5} count distribution and the resulting uncertainty in determining model parameters.

As can be seen in the figure, the IS\textsubscript{5} count distribution in proteobacterial cells is strongly L-shaped. An overwhelming majority of those genomes, namely 92.7\%, does not contain any copies, a small fraction of genomes contains up to 10 or 15 copies, and only few genomes contain more than 15 copies, although there are proteobacterial genomes with higher IS\textsubscript{5} counts: only the \textit{Pseudomonas syringae} tomato DC3000 genome and all three \textit{Xanthomonas oryzae} genomes in our data set contain more than 40 copies of IS\textsubscript{5}, where \textit{Xanthomonas oryzae} MAFF 311018 has the highest count of 54 IS\textsubscript{5} copies.

The HGT rate has to be larger than the fitness cost of IS\textsubscript{5} to reach the observed IS\textsubscript{5} count distribution in equilibrium

We set the replicative transposition rate $u$ and the excision rate $e$ to $(u, e) = (10^{-7}, 10^{-10})$, which is in the range of values provided in table 1. Note that table 1 reports only the conservative transposition rate, and we assume that the replicative transposition rate is a few orders of magnitude smaller \cite{Tavakoli and Derbyshire, 2001}. In the appendix, we present analogous results using a different set $(u, e) = (10^{-9}, 10^{-11})$ of parameters.

First, for all 1000 random samples of size 180 from 525 proteobacterial genomes, we compute the maximum likelihood estimates of the fitness effect $s$ and the HGT parameter $H$ for all 9 possible combinations of the fitness effect exponent $\alpha$ and the replicative transposition exponent $\gamma$. To identify in each sample those models that do not fit the
data significantly worse than the best model of the sample, we follow an argument of Sawyer et al. [Sawyer et al., 1987] and take in each sample the model with the highest log-likelihood as a proxy for the model with free (and continuous) parameters $\alpha$ and $\gamma$. As a consequence, all our models with specific, fixed $\alpha$ and $\gamma$ become nested within this proxy model that is considered to have two additional degrees of freedom. We can then apply the likelihood-ratio test in each sample to compare the proxy model with all other models, using a $\chi^2$ distribution with two degrees of freedom. Therefore, on a 5% significance level, models whose log-likelihood is not more than $\chi^2_{0.05,2}/2 = 3.0$ lower than the log-likelihood of the best model of the sample, fit observed data not significantly worse than this best model. Applied to our data, we find that the exponent combinations $(\alpha, \gamma) = (0, 1)$ and $(\alpha, \gamma) = (1, 1)$ lead to the best fit in 359 and 346 of the 1000 samples, respectively. But based on our criterium for the log-likelihood described above, we find that the exponent combinations $(\alpha, \gamma) = (0, 1)$ (in all 1000 samples), $(\alpha, \gamma) = (2, 2)$ (in 990 samples), $(\alpha, \gamma) = (1, 2)$ (in 957 samples), and $(\alpha, \gamma) = (1, 1)$ (in 951 samples) lead in over 90% of all samples to fits that are not significantly worse than the best fit in each sample.

These findings for $\gamma$ suggest that the transposition rate per IS5 copy is not downregulated with increasing IS5 count per genome. The fitness exponent parameter $\alpha$ does not show a clear distribution pattern, i.e. all its possible values (0, 1, and 2) can lead in over 90% of all samples to a fit that is not significantly worse than the best fit in each sample. Our data does therefore not allow to make a definitive statement about possible interactions between IS5 copies in influencing the fitness of a host cell.

Based on the maximum likelihood estimates of the fitness effect, $\hat{s}$, and of the HGT parameter, $\hat{H}$, we can first compute the total population size $Z = 1 + \frac{\hat{S}^2}{r}$ in equilibrium and then the maximum likelihood estimate of the HGT rate, $\hat{h} = \frac{\hat{H}}{S_{\beta}(q)Z}$, which depends on our choice of the HGT exponent $\beta$. Table 2 shows for the four pairs of $(\alpha, \gamma)$ that are most frequently not significantly worse than the best fitting pair in each sample the quartiles of the maximum likelihood estimates of $s$ and of $h$ for all choices of $\beta \in \{0, 1, 2\}$.

We show only those HGT rates which lead to a stable equilibrium that can be reached by starting with a small proportion of infected cells (between $10^{-9}$ and $10^{-3}$) carrying one copy of IS5.  

*insert table 2 here*
Table 2 shows that \( \hat{s} < 0 \) for IS5, i.e. that IS5 is generally detrimental. The table also shows that \( \beta \leq \alpha \) is needed to reach the equilibrium, starting with a small proportion of infected cells. In that case, \( \hat{h} \geq |\hat{s}| \). If, on the other hand, \( \beta > \alpha \), then \( \hat{h} < |\hat{s}| \). The IS5 infection, starting with cells carrying one copy of IS5 only, will then die out, because HGT is not strong enough to overcome the negative fitness effect caused by even only one IS5 copy per genome. Therefore, for an IS5 infection to spread and persist, the increase in the infectiousness of a cell with increasing IS5 count must be smaller than the simultaneous increase in the total fitness cost.

Above, we found four different exponent pairs \( (\alpha, \beta) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\} \) that fit observed data in more than 900 of 1000 samples not significantly worse than the best fitting pair of each sample. Figure 3 shows for each of these four pairs an example of the predicted equilibrium distribution based on the maximum likelihood estimates \( \hat{s} \) and \( \hat{H} \), together with an observed IS5 count distribution based on a sample that led to the best model fit with the chosen pair \( (\alpha, \gamma) \) and the maximum likelihood estimates \( \hat{s} \) and \( \hat{H} \). The four pairs \( (\alpha, \gamma) \) cover all possible fitness effect exponents \( \alpha \in \{0, 1, 2\} \), and they lead to a wide range of the estimated fitness effect \( \hat{s} \) (compare with table 2). We truncate the computed distribution at \( l = 60 \) copies of IS5 per genome. The bin with 60 copies per genome therefore represents all genomes with at least 60 copies of IS5 in the computed distribution (the highest IS5 count in all proteobacterial genomes is 54 and therefore well below \( l = 60 \)).

A conspicuous feature of the predicted IS count distributions is the sharp upward spike at the highest IS count. It stems from the truncation we imposed at \( l = 60 \) copies of IS5 per genome. In a model with no upper bound for the IS5 count per genome, the distribution would drop monotonously. We have confirmed this by using higher IS5 count limits \( l \) and by observing that the spike in the highest IS5 count then gets smaller when we again apply the maximum likelihood method (results not shown).

To get an estimate of the time needed to approximately reach the equilibrium distribution, we compute the population dynamics of an IS5 infection over time, again for the four exponent pairs \( (\alpha, \beta) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\} \) already used above, and with
the corresponding maximum likelihood estimates of $s$ and $h_\beta = h_0$. We choose to focus on $\beta = 0$, because HGT is tightly regulated and depends on several internal and external factors [Dröge et al., 1999], so that the infectiousness of a cell probably depends only very weakly on the cell genome’s IS count, if at all. Our choice of the base population growth rate $r = 1$ means that one time unit corresponds to the doubling time during the early exponential growth phase of a cell population. We identify this doubling time with one cell generation and set one cell generation to one day [Gibbons and Kapsimalis, 1967, Savageau, 1983] for the purpose of this analysis. Our computations then show, on the one hand, that the time to reach 90% of the final prevalence of infected cells is very long if we start with an initial prevalence of $10^{-6}$ infected cells. It lies between $1.8 \cdot 10^7$ years for $(\alpha, \gamma) = (1, 2)$ and $(\hat{s}, \hat{h}_0) = (-5.6 \cdot 10^{-8}, 1.1 \cdot 10^{-7})$ and $1.8 \cdot 10^{10}$ years for $(\alpha, \gamma) = (0, 1)$ and $(\hat{s}, \hat{h}_0) = (-1.1 \cdot 10^{-7}, 1.1 \cdot 10^{-7})$. On the other hand, the predicted time needed for the population of infected cells only to approximately reach its final IS count distribution is much shorter. It lies between about 7’100 years for $(\alpha, \gamma) = (1, 2)$ and 33’500 years for $(\alpha, \gamma) = (0, 1)$. In the latter computation, we numerically solve the equation \( \frac{1}{2} \sum_{j=1}^{l} |Z_j(t)/Z_{\text{inf}}(t) - Z_j^*/Z_{\text{inf}}^*| = 0.1 \) for the time $t$, where $Z_{\text{inf}} = \sum_{j=1}^{l} Z_j$ and an asterisk (*) indicates the final normalized population densities. We then also compute the dynamics of the host population size over time during an infection with each of the four exponent pairs $(\alpha, \gamma)$, using the same maximum likelihood estimates of $\hat{s}$ and $\hat{h}_0$ as in the preceding paragraph. In all cases, the relative reduction in the normalized population density caused by the infection is negligible, between $4.7 \cdot 10^{-9}$ for $(\alpha, \gamma) = (1, 1)$ and $8.2 \cdot 10^{-9}$ for $(\alpha, \gamma) = (1, 2)$. This is expected, because the fitness cost of IS5 is generally small, as our computations show.

The maximum likelihood estimates of the HGT rate and of the fitness effect are tightly connected

Using 1000 random samples of 180 out of 525 proteobacterial genomes, one per genus, gives an impression of the uncertainty about the real IS5 count distribution and the resulting uncertainty in determining exponent pairs $(\alpha, \gamma)$, fitness effect $s$, and HGT rate $h$. To get information about the variation in the MLEs of $s$ and $h$ due to our model, and to gain some insight into the relationship between $h$ and $s$, we compute confidence intervals for
these two parameters. We accomplish this by using a bootstrap with 1000 resamplings of
the four IS\textsuperscript{5} count distributions for \((\alpha, \beta) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}\) that we have used
before and shown in figure 3. Figure 4 shows the corresponding, bootstrapped values of
\(\hat{h}_0\) versus \(\hat{s}\). We have again chosen to show only the graphs for \(\beta = 0\), as in the preceding
subsection. And we have marked the pairs \((\hat{s}, \hat{h}_0)\) in the 95\% confidence interval of \(s\) with
black dots, while pairs outside the confidence interval are marked with gray dots.

\[\text{insert figure 4 here}\]

The 1000 bootstrapped pairs of \((\hat{s}, \hat{h}_0)\) in figure 4 show an almost perfectly functional
dependence between \(\hat{h}_0\) and \(\hat{s}\). To concisely describe this functional dependence, we have
plotted the graph of the best fit of the shifted power function \(\hat{h}_0 = a(-\hat{s})^b + c\). As can be
seen, the fit is very good. The functional dependence between \(\hat{h}_0\) and \(\hat{s}\) is almost linear
if \(\beta = \alpha\). We can understand this linear dependence by observing that from the first
equation in the differential equation system (2) we get in equilibrium

\[h = \frac{ep_1 - sp_0S_{\alpha}(p)}{p_0ZS_{\beta}(p)} \approx -\frac{S_{\alpha}(p)}{S_{\beta}(p)}s.\]  

(5)

Our model therefore suggests that the maximum likelihood estimate of the HGT rate
depends very sensitively and almost exclusively on the maximum likelihood estimate of
the fitness effect (and vice versa). This highlights the crucial role the HGT rate plays in
surmounting the fitness cost of IS\textsuperscript{5} and allowing IS\textsuperscript{5} to persist in a host cell population.

Discussion

While an IS that provides a benefit to its host can rise to fixation through natural se-
lection [Hall, 1999, Schneider and Lenski, 2004], the outcome of an infection with purely
detrimental ISs is less clear. We have shown in an earlier paper that regardless of whether
ISs are moderately beneficial or detrimental, the chances of a successful IS infection are
small [Bichsel et al., 2010]. Here we are interested in the longer-term fate of an IS infec-
tion. Specifically, we investigate whether a purely detrimental IS can persist and reach the
IS\textsuperscript{5} count distribution in natural, proteobacterial host cell populations, where IS\textsuperscript{5} mainly
occurs. We are also interested in the fitness effect \(s\) and in the HGT rate \(h\) needed to
reach this IS5 count distribution.

We analyze 525 fully sequenced genomes from 180 proteobacterial genera and use the IS5 count distribution in these genomes to compute the maximum likelihood estimates of the fitness effect and of the HGT rate. The sequenced genomes stem from various proteobacterial cell populations all over the world and do not constitute a genome sample from a single population. At first sight, it is therefore not clear that we can compare the IS5 count distribution in the sequenced genomes with the IS count distribution our model of a single population predicts. However, we note that a very similar, L-shaped IS5 count distribution has also been observed in the ECOR collection of 71 strains of *Escherichia coli* [Sawyer et al., 1987], which is a less heterogeneous sample that covers a smaller taxonomic range than the proteobacterial genomes in our data set. This observations, together with the fact that the L-shaped IS count distribution can be observed in other IS families [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007], motivates our assumption that this distribution does not depend on a specific IS and on the taxonomic scale. We thus assume that the same distribution would also exist in other ISs and on the smallest taxonomic scale, that of a cell population.

Another objection to our approach might consist in observing that we use IS5 count data from phylogenetically related genomes to conduct a maximum likelihood analysis which assumes independence between observations. There is no denying that the genomes in our data set are tied together phylogenetically and that their IS5 counts are therefore not strictly independent of each other. Nevertheless, we can reduce this dependence by choosing only one genome per genus for the maximum likelihood analysis. To still get a proper basis for our conclusions, we generated 1000 sample data sets, each containing one genome per genus and repeated the maximum likelihood analysis for each sample data set.

It might be argued that the IS5 count distribution in our data set is influenced by the fact that many ISs show DNA target specificities of varying degrees [Chandler and Mahillon, 2002]: IS1 prefers AT-rich regions, and IS4 is known to preferably insert into DNA sequences of the form AAA–N$_{15–20}$–TTT [Zerbib et al., 1985, Mayaux et al., 1984]. And in fact, IS5 does show some preference for the target sequence CTAG. However, target specificity is probably not strong enough to limit the IS5 count distribution noticeably in the IS count...
range on which we base our computations (0–60 copies of IS\textit{5} per genome). This is supported by the observation that some proteobacterial genomes contain more than 40 copies of IS\textit{5}, although most of the genomes have much lower IS\textit{5} counts.

We now discuss the main points of this study.

**Purely detrimental ISs can persist if the HGT rate is larger than the fitness cost of an IS**

The L-shaped IS\textit{5} count distribution in 525 sequenced proteobacterial genomes (and assumedly in natural host cell populations) suggests that IS\textit{5} (and probably all ISs with similar IS count distribution) is generally detrimental to its hosts. Our results support this suggestion and show that even purely detrimental ISs may persist and reach an IS count distribution similar to the one observed in IS\textit{5} in sequenced genomes, provided that the HGT rate is larger than the fitness cost induced by one IS in the genome of an infected cell. This is in agreement with our earlier result based on a stochastic infection model. The HGT rate in turn is larger than the fitness cost of one IS only if the possible increase in the infectiousness of a cell is smaller than the increase in the fitness cost with an increasing IS count. A small increase of the infectiousness with an increasing IS count is consistent with earlier observations that HGT is tightly regulated and depends on many different factors [Dröge et al., 1999], so that the influence of the IS count on the HGT rate, and therefore on the infectiousness of a cell, is probably small, if it exists at all.

The observed IS\textit{5} count distribution suggests that the replicative transposition rate of IS\textit{5} is not down-regulated

Our model shows in 797 of 1000 samples best agreement with the IS\textit{5} count distribution in sequenced proteobacterial genomes if the replicative transposition rate increases linearly with the IS\textit{5} count per genome, i.e. if replicative transposition is not regulated and copies of IS\textit{5} transpose independently. This is in agreement with results published by Sawyer et al. [Sawyer et al., 1987]. Using branching processes to model the IS count distribution in the ECOR collection of 71 natural isolates of \textit{Escherichia coli}, these authors report that a linear dependence of the replicative transposition rate on the IS count agrees best with the available data in the collection for IS\textit{5}. Sawyer et al. use for their analysis the ECOR
collection, which is a smaller albeit more homogeneous dataset than our collection of 525 sequenced proteobacterial genomes. Besides using a larger dataset, our analysis is based on an ordinary differential equation model that allows for interactions between cells and for density-dependent population growth and infection, which makes it more suitable to analyse the long-term fate of an IS infection than the branching process model used by Sawyer et al.

**IS5 might be effectively neutral to its hosts**

Our model predicts a fitness effect in the range $\hat{s} \in [-10^{-7}, -10^{-9}]$ for IS5 (see table 2). Considering that the effective population size of typical prokaryotes is of the order of $N_e \approx 10^8$ [Lynch, 2007, p. 92], IS5 might therefore be effectively neutral or only slightly detrimental to its hosts. Hence, HGT is probably strong enough to enable IS5 to persist and spread in a host cell population (see table 1). At the same time, our model predicts an unrealistically long time for IS5 to approximately reach the final prevalence of infected cells, while the predicted time to approximately reach the IS5 count distribution in infected cells only is much shorter. It therefore seems that the timescale of the infection process may be much larger than the timescale of the process that leads to an equilibrium distribution in the population of infected cells. While the former timescale is determined by the antagonistic actions of HGT and the fitness cost of one copy of IS5, the latter timescale is determined mainly by replicative transposition and the fitness cost of varying numbers of IS5 copies. This observation of different timescales leads us to suggest that IS5 may have been at least occasionally and temporarily beneficial to its host cells, which can accelerate its spreading through single populations and through populations all over the world.

**Authors contributions**

All three authors contributed to the design of the study. MB carried out the model analysis. ADB contributed to the mathematical analysis of the ordinary differential equation model. AW contributed to the biological interpretation of the analytical results. MB and AW contributed to writing the manuscript.
Acknowledgements

MB and AW would like to acknowledge support from Swiss National Science Foundation grants 315200-116814 and 315200-119697, as well as from the YeastX grant of SystemsX.ch.
Appendix

Results for other replicative transposition and excision rates

We have repeated our calculations for another combination of the replicative transposition rate $u$ and the excision rate $e$, this time at the lower end of the rate range described in table 1: $(u, e) = (10^{-9}, 10^{-11})$. Because the effect of the excision rate is small, and because the effect of an IS5 infection on the normalized population density $Z$ is again negligible, the fitness effect $s$ and the HGT rate $h$ scale almost linearly with the assumed transposition rate $u$ (see the differential equation system (2, 3)). This is exactly what can be observed.

Table 3 shows for the four pairs of $(\alpha, \gamma)$ that are most frequently not significantly worse than the best fitting pair in each sample the quartiles of the maximum likelihood estimates of $s$ and of $h$ for all choices of $\beta \in \{0, 1, 2\}$. We show only those HGT rates which lead to a stable equilibrium that can be reached by starting with a small proportion of infected cells (between $10^{-9}$ and $10^{-3}$) carrying one copy of IS5.

As can be seen when compared with table 2 in the main text, the quartiles of the maximum likelihood estimates of $s$ and $h$ scale almost perfectly linearly with the new choice for the replicative transposition rate $u$.

We draw the same conclusions as in the main text: IS5 seems to be effectively neutral (even more so for this parameter combination of $u$ and $e$), and HGT ist most probably strong enough to overcome the fitness cost caused by a copy of IS5 in the host cell genome.
Figures

\[ r(1 - Z)Z_0 \quad [r(1 - Z) + s]Z_1 \quad [r(1 - Z) + l^\alpha s]Z_l \]

\[ h Z_0 \sum_{k=1}^{l} k^\beta Z_k \quad uZ_1 \quad (l - 1)\gamma uZ_{l-1} \]

\[ eZ_1 \quad 2eZ_2 \quad leZ_l \]

Figure 1: Model design. \( Z_k = D_k/K \) is the normalized density of cells with \( k \in \{0, \ldots, l\} \) ISs, where \( D_k \) is the density of cells with \( k \) ISs, and \( K \) is the carrying capacity; \( r = 1 \) is the base growth rate per uninfected cell; \( s \) is the fitness effect of one IS; \( u \) is the base replicative transposition rate of one IS; \( e \) is the excision rate per IS; \( h \) is the HGT rate; \( l = 60 \) is the maximal IS count per genome; \( \alpha, \beta, \gamma \in \{0, 1, 2\} \) are power function exponents that control the increase of the fitness effect, of the HGT rate, and of the replicative transposition rate with increasing IS count per cell. All rates are per time unit. Because \( r = 1 \), one time unit corresponds to the doubling time during the exponential cell growth phase. Solid lines indicate a change in the total cell density, and dashed lines indicate a change only in the normalized density distribution of the cells with different IS counts in their genome.

Figure 2: Mean IS count distribution of IS5 over 1000 random samples. Each sample consists of 180 genomes, one randomly chosen genome per proteobacterial genus. In total, the proteobacteria in our data set consist of 525 fully sequenced genomes in 180 genera. Note the logarithmic scale on the vertical axis.
Figure 3: Observed (large, gray dots) and predicted (small, black dots connected by a solid line) IS\textsuperscript{5} count distributions for (\(\alpha, \gamma\)) \(\in\) \{(0, 1), (1, 1), (1, 2), (2, 2)\}. In the predicted IS\textsuperscript{5} count distribution, the IS\textsuperscript{5} count per genome has been limited to \(l = 60\) copies of IS\textsuperscript{5}. Note the logarithmic scale on the vertical axis.
Figure 4: Bootstrapped pairs of \((\hat{s}, \hat{h}_0)\) for \((\alpha, \gamma) = \{(0, 1), (1, 1), (1, 2), (2, 2)\}\), based on 1000 resamplings of four IS\(^5\) count distributions in proteobacteria. Black dots lie inside and gray dots lie outside the 95% confidence interval of \(s\). The values of \((\hat{s}, \hat{h}_0)\) for the original IS\(^5\) count distribution are marked by large, black dots. The graphs of the shifted power function approximations \(\hat{h}_0 = a \cdot (-\hat{s})^b + c\) are shown as thin lines. The parameters \((a, b, c)\) are \((1.00, 1.00, 2.1 \cdot 10^{-11})\) for \((\alpha, \gamma) = (0, 1)\), \((0.063, 0.818, 2.5 \cdot 10^{-8})\) for \((\alpha, \gamma) = (1, 1)\), \((1.125, 1.009, 6.0 \cdot 10^{-8})\) for \((\alpha, \gamma) = (1, 2)\), and \((0.148, 0.876, 6.2 \cdot 10^{-8})\) for \((\alpha, \gamma) = (2, 2)\).
Tables

<table>
<thead>
<tr>
<th>Event</th>
<th>Rates</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transposition</td>
<td>Conservative</td>
<td>$10^{-7} - 10^{-4}$ [Kleckner, 1989],</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Chandler and Mahillon, 2002]</td>
</tr>
<tr>
<td>Excision</td>
<td></td>
<td>$10^{-10}$ [Kleckner, 1989]</td>
</tr>
<tr>
<td>HGT Transformation</td>
<td>$10^{-6} - 10^{-3}$ [Williams et al., 1996]</td>
<td></td>
</tr>
<tr>
<td>Transduction</td>
<td></td>
<td>$10^{-8}$ [Jiang and Paul, 1998]</td>
</tr>
<tr>
<td>Conjugation</td>
<td></td>
<td>$10^{-6} - 10^{-5}$ [Dahlberg et al., 1998]</td>
</tr>
</tbody>
</table>

Table 1: Transposition, excision, and HGT rates reported by different authors. Rates have been converted into events per cell or IS and generation.

<table>
<thead>
<tr>
<th>$(\alpha, \gamma)$</th>
<th>Quart.</th>
<th>$\hat{s}$</th>
<th>$\hat{h}_0$</th>
<th>$\hat{h}_1$</th>
<th>$\hat{h}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 1)</td>
<td>Q1</td>
<td>$-1.6 \cdot 10^{-7}$</td>
<td>$1.1 \cdot 10^{-7}$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>$-1.3 \cdot 10^{-7}$</td>
<td>$1.3 \cdot 10^{-7}$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>$-1.1 \cdot 10^{-7}$</td>
<td>$1.6 \cdot 10^{-7}$</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>(1, 1)</td>
<td>Q1</td>
<td>$-2.8 \cdot 10^{-8}$</td>
<td>$3.8 \cdot 10^{-8}$</td>
<td>$7.3 \cdot 10^{-9}$</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>$-1.8 \cdot 10^{-8}$</td>
<td>$5.5 \cdot 10^{-8}$</td>
<td>$1.8 \cdot 10^{-8}$</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>$-7.3 \cdot 10^{-9}$</td>
<td>$6.7 \cdot 10^{-8}$</td>
<td>$2.8 \cdot 10^{-8}$</td>
<td>–</td>
</tr>
<tr>
<td>(1, 2)</td>
<td>Q1</td>
<td>$-2.9 \cdot 10^{-8}$</td>
<td>$7.7 \cdot 10^{-8}$</td>
<td>$1.9 \cdot 10^{-8}$</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>$-2.3 \cdot 10^{-8}$</td>
<td>$8.2 \cdot 10^{-8}$</td>
<td>$2.3 \cdot 10^{-8}$</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>$-1.9 \cdot 10^{-8}$</td>
<td>$8.8 \cdot 10^{-8}$</td>
<td>$2.9 \cdot 10^{-8}$</td>
<td>–</td>
</tr>
<tr>
<td>(2, 2)</td>
<td>Q1</td>
<td>$-3.1 \cdot 10^{-9}$</td>
<td>$6.3 \cdot 10^{-8}$</td>
<td>$1.8 \cdot 10^{-8}$</td>
<td>$7.9 \cdot 10^{-10}$</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>$-1.6 \cdot 10^{-9}$</td>
<td>$6.5 \cdot 10^{-8}$</td>
<td>$2.3 \cdot 10^{-8}$</td>
<td>$1.6 \cdot 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>$-7.9 \cdot 10^{-10}$</td>
<td>$6.8 \cdot 10^{-8}$</td>
<td>$2.9 \cdot 10^{-8}$</td>
<td>$3.1 \cdot 10^{-9}$</td>
</tr>
</tbody>
</table>

Table 2: For a replicative transposition rate $u = 10^{-7}$ and an excision rate $e = 10^{-10}$, the table shows the four most frequent exponents $(\alpha, \gamma) \in \{0, 1, 2\}^2$ that lead to model fits of the IS5 count distribution that are not significantly worse than the best fit. For each pair $(\alpha, \gamma)$, the quartiles of the maximum likelihood estimates of the fitness effect $s$ and of the HGT rate $h_\beta$ for different scaling exponents $\beta \in \{0, 1, 2\}$ of the HGT rate are reported. Only HGT rates that lead to stable equilibria are shown. Observe that Q1 in $\hat{s}$ corresponds to Q3 in $\hat{h}_\beta$ and vice versa.
Table 3: For a replicative transposition rate $u = 10^{-9}$ and an excision rate $e = 10^{-11}$, the table shows the four most frequent exponents $(\alpha, \gamma) \in \{0, 1, 2\}^2$ that lead to model fits of the IS5 count distribution that are not significantly worse than the best fit. For each pair $(\alpha, \gamma)$, the quartiles of the maximum likelihood estimates of the fitness effect $s$ and of the HGT rate $h_\beta$ for different scaling exponents $\beta \in \{0, 1, 2\}$ of the HGT rate are reported. Only HGT rates that lead to stable equilibria are shown. Observe that Q1 in $\hat{s}$ corresponds to Q3 in $\hat{h}_\beta$ and vice versa.

<table>
<thead>
<tr>
<th>$(\alpha, \gamma)$ Quart.</th>
<th>$\hat{s}$</th>
<th>$\hat{h}_0$</th>
<th>$\hat{h}_1$</th>
<th>$\hat{h}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 1) Q1</td>
<td>$-1.6 \cdot 10^{-9}$</td>
<td>1.1 $\cdot 10^{-9}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>M</td>
<td>$-1.3 \cdot 10^{-9}$</td>
<td>1.3 $\cdot 10^{-9}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Q3</td>
<td>$-1.1 \cdot 10^{-9}$</td>
<td>1.6 $\cdot 10^{-9}$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(1, 1) Q1</td>
<td>$-2.8 \cdot 10^{-10}$</td>
<td>3.8 $\cdot 10^{-10}$</td>
<td>7.3 $\cdot 10^{-11}$</td>
<td>-</td>
</tr>
<tr>
<td>M</td>
<td>$-1.8 \cdot 10^{-10}$</td>
<td>5.5 $\cdot 10^{-10}$</td>
<td>1.8 $\cdot 10^{-10}$</td>
<td>-</td>
</tr>
<tr>
<td>Q3</td>
<td>$-7.3 \cdot 10^{-11}$</td>
<td>6.7 $\cdot 10^{-10}$</td>
<td>2.8 $\cdot 10^{-10}$</td>
<td>-</td>
</tr>
<tr>
<td>(1, 2) Q1</td>
<td>$-2.9 \cdot 10^{-10}$</td>
<td>7.8 $\cdot 10^{-10}$</td>
<td>1.9 $\cdot 10^{-10}$</td>
<td>-</td>
</tr>
<tr>
<td>M</td>
<td>$-2.3 \cdot 10^{-10}$</td>
<td>8.2 $\cdot 10^{-10}$</td>
<td>2.3 $\cdot 10^{-10}$</td>
<td>-</td>
</tr>
<tr>
<td>Q3</td>
<td>$-1.9 \cdot 10^{-10}$</td>
<td>8.8 $\cdot 10^{-10}$</td>
<td>2.9 $\cdot 10^{-10}$</td>
<td>-</td>
</tr>
<tr>
<td>(2, 2) Q1</td>
<td>$-3.1 \cdot 10^{-11}$</td>
<td>6.4 $\cdot 10^{-10}$</td>
<td>1.8 $\cdot 10^{-10}$</td>
<td>8.0 $\cdot 10^{-12}$</td>
</tr>
<tr>
<td>M</td>
<td>$-1.6 \cdot 10^{-11}$</td>
<td>6.6 $\cdot 10^{-10}$</td>
<td>2.4 $\cdot 10^{-10}$</td>
<td>1.6 $\cdot 10^{-11}$</td>
</tr>
<tr>
<td>Q3</td>
<td>$-7.9 \cdot 10^{-12}$</td>
<td>6.8 $\cdot 10^{-10}$</td>
<td>2.9 $\cdot 10^{-10}$</td>
<td>3.1 $\cdot 10^{-11}$</td>
</tr>
</tbody>
</table>

References


detection of high levels of horizontal plasmid transfer in marine bacterial communities.
*Applied and Environmental Microbiology*, 64(7):2670–2675.


transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm

the Bootstrap*. Chapman & Hall/CRC.

American Society for Microbiology, Washington, D.C.

the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice.


D.C.

[Lan and Reeves, 2002] Lan, R. and Reeves, P. R. (2002). *Escherichia coli* in disguise:
Estimating the Fitness Effect of the Insertion Sequence IS5


Estimating the Fitness Effect of the Insertion Sequence IS5


