J. Math. Biol. (2013) 66:95–114
DOI 10.1007/s00285-012-0504-2

Mathematical Biology

Estimating the fitness effect of an insertion sequence

Manuel Bichsel · A. D. Barbour · Andreas Wagner

Received: 10 March 2011 / Revised: 30 October 2011 / Published online: 18 January 2012

Abstract Since its discovery, mobile DNA has fascinated researchers. In particular, many researchers have debated why insertion sequences persist in prokaryote genomes and populations. While some authors think that insertion sequences persist only because of occasional beneficial effects they have on their hosts, others argue that horizontal gene transfer is strong enough to overcome their generally detrimental effects. 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 per cell. Based on our model and the distribution of IS*5*, an insertion sequence for which sufficient data is available in 525 fully sequenced proteobacterial genomes, we show that the fitness cost of insertion sequences is so small that they are effectively neutral or only slightly detrimental. We also show that an insertion sequence 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*5* is unrealistically long for the fitness cost and horizontal gene transfer rate that we computed. Occasional beneficial effects may thus have played an important role in the fast spreading of insertion sequences like IS*5*.

Keywords Mobile DNA · Insertion sequence · Fitness effect · Horizontal gene transfer · Ordinary differential equation system · Maximum likelihood

M. Bichsel (✉) · A. Wagner
Institute of Evolutionary Biology and Environmental Studies,
University of Zürich, Zuerich, Switzerland
e-mail: manuel.bichsel@ieu.uzh.ch

A. D. Barbour
Institute of Mathematics, University of Zürich, Zuerich, Switzerland

Mathematics Subject Classification (2000) 92B15 General biostatistics · 92D25 Population dynamics (general)

1 Introduction

Bacterial insertion sequences (ISs) are the simplest form of autonomous mobile DNA. They are short (800–2,500 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. IS*1*A, IS*5*, IS*630*. 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, IS*5*, belongs to a rather heterogeneous family of ISs that is widely distributed among bacteria and archaea.

IS*5* 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 form 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).

Table 1 Transposition, excision, and HGT rates reported by different authors

Event		Rates	Sources
Transposition	Conservative	10^{-7}–10^{-4}	Kleckner (1989), Chandler and Mahillon (2002)
Excision		10^{-10}	Kleckner (1989)
HGT	Transformation	10^{-6}–10^{-3}	Williams et al. (1996)
	Transduction	10^{-8}	Jiang and Paul (1998)
	Conjugation	10^{-6}–10^{-5}	Dahlberg et al. (1998)

Rates have been converted into events per cell or IS and generation

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. Those infections that do persist, take a very long time to reach noticeable population sizes. 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 obtain the IS count distribution we observe in bacterial genomes. In doing so, we focus on the largely proteobacterial insertion sequence IS*5*, the only IS for which sufficient data is available. Because very similar IS count distributions have been observed in many other ISs (Sawyer et al. 1987; Wagner 2006; Touchon and Rocha 2007), we presume that our results are at least qualitatively comparable to those that would be obtained for other ISs.

2 Data, model, and methods

2.1 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 1,447 genome sequences (only chromosomes, no plasmids) for these 20 representative IS sequences. For later analysis, we needed independent IS count observations. We were therefore interested in ISs that occur in many genera. Of all 20 ISs, only 3 occur in more than 10 different genera. And of these 3 ISs, only IS*5* 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 IS*5*. It turned out that IS*5* (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 IS*5* 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

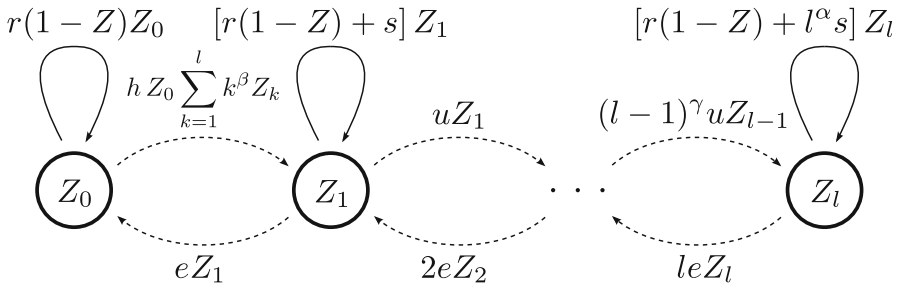

Fig. 1 Model design. $Z_k = D_k/K$ is the normalized density of cells with $k \in \{0, \ldots, l\}$ ISs, where D_k is the density of cells with k ISs, and K is the carrying capacity; $r = 1$ is the base growth rate per uninfected cell; s is the base fitness effect of one IS; u is the base replicative 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

Escherichia because of their well-known close phylogenetic relationship (Lan and Reeves 2002).

2.2 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). Furthermore, we show in the results section that no genome in our data set contains more than 60 copies of IS*5*, 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 base 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 α, β and γ, 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 Fig. 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).

2.3 Model analysis

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

$$
\begin{aligned}
\dot{Z}_0 &= r(1-Z)Z_0 - h\,Z_0 \sum\nolimits_{k=1}^{l} k^{\beta} Z_k + eZ_1 \\
\dot{Z}_1 &= [r(1-Z)+s]\,Z_1 + h\,Z_0 \sum\nolimits_{k=1}^{l} k^{\beta} Z_k - uZ_1 - eZ_1 + 2eZ_2 \\
\dot{Z}_2 &= \left[r(1-Z)+2^{\alpha}s\right] Z_2 + uZ_1 - 2^{\gamma}uZ_2 - 2eZ_2 + 3eZ_3 \\
&\vdots \\
\dot{Z}_j &= \left[r(1-Z)+j^{\alpha}s\right] Z_j + (j-1)^{\gamma}uZ_{j-1} - j^{\gamma}uZ_j - jeZ_j + (j+1)eZ_{j+1} \\
&\vdots \\
\dot{Z}_l &= \left[r(1-Z)+l^{\alpha}s\right] Z_l + (l-1)^{\gamma}uZ_{l-1} - leZ_l
\end{aligned}
\tag{1}
$$

This system has two obvious equilibrium solutions: the first one is $Z_0 = Z_1 = \cdots = Z_l = 0$, i.e. population extinction, and the second one is $Z_0 = 1$ and $Z_1 = \cdots = Z_l = 0$, 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,

$$\dot{p_k} = \frac{\dot{Z_k}}{Z} - \frac{Z_k \cdot \dot{Z}}{Z^2} = \frac{1}{Z}\dot{Z_k} - p_k \frac{1}{Z}\sum_{j=0}^{l} \dot{Z_j} = \frac{1}{Z}\dot{Z_k} - p_k \left(r(1-Z) + s\sum_{j=1}^{l} j^\alpha p_j \right),$$

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

$$\begin{aligned}
\dot{p_0} &= -h\, p_0\, Z \textstyle\sum_{k=1}^{l} k^\beta p_k + e p_1 - s\, p_0 \sum_{k=1}^{l} k^\alpha p_k \\
\dot{p_1} &= s p_1 + h\, p_0\, Z \textstyle\sum_{k=1}^{l} k^\beta p_k - u p_1 - e p_1 + 2e p_2 - s\, p_1 \sum_{k=1}^{l} k^\alpha p_k \\
\dot{p_2} &= 2^\alpha s p_2 + u p_1 - 2^\gamma u p_2 - 2e p_2 + 3e p_3 - s\, p_2 \textstyle\sum_{k=1}^{l} k^\alpha p_k \\
&\vdots \\
\dot{p_j} &= j^\alpha s p_j + (j-1)^\gamma u p_{j-1} - j^\gamma u p_j - je p_j + (j+1) e p_{j+1} - s\, p_j \textstyle\sum_{k=1}^{l} k^\alpha p_k \\
&\vdots \\
\dot{p_l} &= l^\alpha s p_l + (l-1)^\gamma u p_{l-1} - le p_l - s\, p_l \textstyle\sum_{k=1}^{l} k^\alpha p_k
\end{aligned} \tag{2}$$

and

$$\dot{Z} = \sum_{k=0}^{l} \dot{Z_k} = r(1-Z)Z + s\, Z \sum_{k=1}^{l} k^\alpha p_k = \left(r(1-Z) + s\sum_{k=1}^{l} k^\alpha p_k \right) Z. \tag{3}$$

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 $(u, e) = \left(10^{-7}, 10^{-10}\right)$, and in the appendix we use $(u, e) = \left(10^{-9}, 10^{-11}\right)$. To solve the system (2, 3), we define $\mathbf{p} = (p_0, \ldots, p_l)^T$, $S_\alpha(\mathbf{p}) = \sum_{k=1}^{l} k^\alpha p_k$, $S_\beta(\mathbf{p}) = \sum_{k=1}^{l} k^\beta p_k$, and the HGT parameter $H(\mathbf{p}, Z) = h\, S_\beta(\mathbf{p})\, Z$. Observe that $H \geq 0$. The differential equations for $p_0, \ldots, p_l$ in (2) can now be written in vector notation as

$$\dot{\mathbf{p}} = \mathbf{M}(s, H(\mathbf{p}, Z)) \cdot \mathbf{p} - s\, S_\alpha(\mathbf{p})\, \mathbf{p} \tag{4}$$

where

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

We get again the IS extinction equilibrium for $\mathbf{p} = \mathbf{e}_0 = (1, 0, \ldots, 0)^T$, because $S_\alpha(\mathbf{e}_0) = S_\beta(\mathbf{e}_0) = H(\mathbf{e}_0, Z) = 0$ for any $Z > 0$, and therefore $\mathbf{M}(s, 0) \cdot \mathbf{e}_0 = \mathbf{0}$, so

that $\mathbf{M}(s, 0) \cdot \mathbf{p} - s\, S_\alpha(\mathbf{p})\, \mathbf{p} = \mathbf{0}$. For all other equilibrium solutions $(\mathbf{p}, Z)$ of (2, 3) that may exist, $H = H(\mathbf{p}, Z)$ and $\lambda = s\, S_\alpha(\mathbf{p})$ must fulfill $\mathbf{M}(s, H) \cdot \mathbf{p} = \lambda\, \mathbf{p}$. We are therefore looking for non-negative eigenvectors of the matrix $\mathbf{M}(s, H)$ for $H > 0$ ($H = 0$ is not interesting, because it implies $Z_1 = \cdots = Z_l = 0$).

$\mathbf{M}(s, H)$ for $H > 0$ is a Metzler–Leontief matrix, i.e. $(\mathbf{M})_{ij} \geq 0$ for $i \neq j$ (Seneta 1981, p. 45). In addition, $\mathbf{M}$ is irreducible. Therefore, for any choice of $H > 0$, there exists an eigenvalue $\tau \in \mathbb{R}$ such that $\tau > \mathrm{Re}(\mu)$ for all other eigenvalues μ of $\mathbf{M}$, and there exists a unique (up to multiples), strictly positive eigenvector $\mathbf{q}$ associated with τ. $\mathbf{q}$ 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 $(1, \ldots, 1) \cdot \mathbf{M}(s, H) \cdot \mathbf{p} = \eta\, (1, \ldots, 1) \cdot \mathbf{p} = \eta$,
2. $(1, \ldots, 1) \cdot \mathbf{M}(s, H) \cdot \mathbf{p} = s\, S_\alpha(\mathbf{p}) = \lambda$ 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 $Z = 1 + \frac{s}{r} S_\alpha(\mathbf{q})$, so that $\dot{Z}(\mathbf{q}) = \left[r(1 - Z) + s\, S_\alpha(\mathbf{q})\right] Z = 0$, 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, β, 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 $\beta \in \{0, 1, 2\}$. These are the required computational steps:

1. Compute the unique eigenvector $\mathbf{q}$ with $\|\mathbf{q}\|_1 = 1$ that corresponds to the (real) eigenvalue τ with the largest real part of the matrix $\mathbf{M}(s, H)$.
2. Set $Z = 1 + \frac{s}{r} S_\alpha(\mathbf{q}) = 1 + \frac{s}{r} \sum_{k=1}^{l} k^\alpha q_k$.
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 $J = \left(\frac{\partial f_i(Z_0, \ldots, Z_l)}{\partial Z_j}\right)_{i,j \in \{0,\ldots,l\}}$ at the specific values of (s, h) and (α, β, γ) used to compute the equilibrium. Here, $f_i(Z_1, \ldots, Z_l)$ 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 (α, β, γ) 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 (α, β, γ) 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 (α, β, γ) and (s, h). We start by defining the likelihood function L. Given the observed IS counts $(c_0, \ldots, c_l)$ and the predicted IS count distribution $(q_0, \ldots, q_l)$ based on the parameters α, γ, s and H, the likelihood function is given by

$$L(\alpha, \gamma, s, H) = q_0^{c_0} \cdot \ldots \cdot q_l^{c_l},$$

and its (natural) logarithm is

$$\ln\left(L(\alpha, \gamma, s, H)\right) = c_0 \cdot \ln(q_0) + \cdots + c_l \cdot \ln(q_l).$$

Because we can only numerically compute the vector of proportions $\mathbf{q}$ based on the parameters α, γ, s, and H, and because we cannot derive $\mathbf{q}$ 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 α and γ 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.

For four specific maximum likelihood estimates $(\hat{s}, \hat{H})$, based on four different exponent pairs (α, γ) we then use the bootstrap method with 1,000 artificially generated data sets 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).

3 Results

3.1 The IS*5* count distribution in proteobacterial cells is L-shaped

Figure 2 shows the IS*5* count distribution based on 525 fully sequenced, proteobacterial genomes. We have generated 1,000 random samples of genomes. Each sample consists of 180 genomes, one randomly chosen genome per proteobacterial genus. We then counted how many genomes per random sample contained 0 ISs, 1–5 ISs, …, 16–20 ISs, or more than 20 ISs. Figure 2 shows the mean number of genomes per IS count bin over all 1,000 samples, together with the 10th and 90th percentile. Averaging over 1,000 random samples provides us with an approximation of the real IS*5* count distribution over proteobacterial genera. Furthermore the 1,000 random samples provide insight into the uncertainty about the real IS*5* count distribution, and about the resulting uncertainty in determining model parameters.

As can be seen in the figure, the IS*5* count distribution in proteobacterial cells is strongly L-shaped. An overwhelming majority of genomes, namely 92.7%, does not contain any IS*5* copies, a small fraction of genomes contains up to 10 or 15 copies,

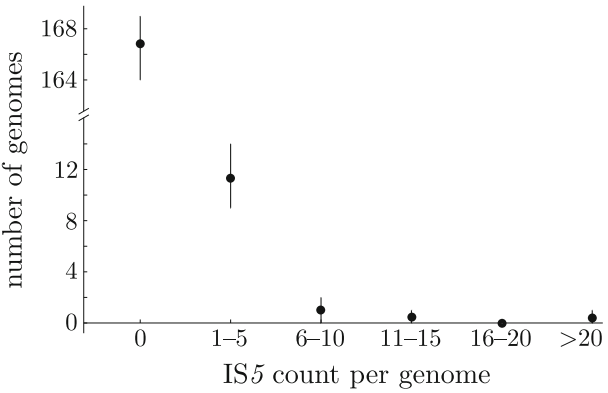

Fig. 2 IS*5* count distribution of 1,000 random samples from 525 proteobacterial genomes, each sample containing 180 genomes. Different IS counts have been collected into bins. *Dots* mark the mean number of genomes in the corresponding bin, and the lower and upper ends of the vertical lines mark the 10th and 90th percentile of the number of genomes, respectively. Note the discontinuous scale on the *vertical axis*

and only few genomes contain more than 15 copies, although there are proteobacterial genomes with higher IS*5* counts. The *Pseudomonas syringae* tomato DC3000 genome and all three *Xanthomonas oryzae* genomes in our data set contain more than 40 copies of IS*5*, where *Xanthomonas oryzae* MAFF 311018 has the highest count of 54 IS*5* copies.

3.2 The HGT rate has to be larger than the fitness cost of IS*5* for an IS infection to reach the observed IS*5* count distribution in equilibrium

We set the replicative transposition rate u and the excision rate e to $(u, e) = \left(10^{-7}, 10^{-10}\right)$, 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 (Tavakoli and Derbyshire 2001). In the appendix, we present analogous results using a different set $(u, e) = \left(10^{-9}, 10^{-11}\right)$ of parameters.

Next, 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 α and the replicative transposition exponent γ. We do so for all 1,000 random samples of size 180 from 525 proteobacterial genomes. To identify in each sample those models that do not fit the data significantly worse than the best model for the sample, we follow an argument of 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 α and γ. As a consequence, all our models with specific, fixed α and γ 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 χ^2 distribution with two degrees of freedom. Therefore, on a 5% significance level, models whose log-likelihood is not by more than $\chi^2_{0.05,2}/2 = 3.0$ units lower than the log-likelihood of the best model of the sample, fit observed data

Table 2 For a replicative transposition rate $u = 10^{-7}$ and an excision rate $e = 10^{-10}$, the table shows the four most frequent exponent pairs $(\alpha, \gamma) \in \{0, 1, 2\}^2$ that lead to model fits of the IS*5* count distribution that are not significantly worse than the best fit

(α, γ)	Quart.	$\hat{s}$	$\hat{h_0}$	$\hat{h_1}$	$\hat{h_2}$
(0, 1)	Q1	$-1.6 \cdot 10^{-7}$	$1.1 \cdot 10^{-7}$	–	–
	Q2	$-1.3 \cdot 10^{-7}$	$1.3 \cdot 10^{-7}$	–	–
	Q3	$-1.1 \cdot 10^{-7}$	$1.6 \cdot 10^{-7}$	–	–
(1, 1)	Q1	$-2.8 \cdot 10^{-8}$	$3.8 \cdot 10^{-8}$	$7.3 \cdot 10^{-9}$	–
	Q2	$-1.8 \cdot 10^{-8}$	$5.5 \cdot 10^{-8}$	$1.8 \cdot 10^{-8}$	–
	Q3	$-7.3 \cdot 10^{-9}$	$6.7 \cdot 10^{-8}$	$2.8 \cdot 10^{-8}$	–
(1, 2)	Q1	$-2.9 \cdot 10^{-8}$	$7.7 \cdot 10^{-8}$	$1.9 \cdot 10^{-8}$	–
	Q2	$-2.3 \cdot 10^{-8}$	$8.2 \cdot 10^{-8}$	$2.3 \cdot 10^{-8}$	–
	Q3	$-1.9 \cdot 10^{-8}$	$8.8 \cdot 10^{-8}$	$2.9 \cdot 10^{-8}$	–
(2, 2)	Q1	$-3.1 \cdot 10^{-9}$	$6.3 \cdot 10^{-8}$	$1.8 \cdot 10^{-8}$	$7.9 \cdot 10^{-10}$
	Q2	$-1.6 \cdot 10^{-9}$	$6.5 \cdot 10^{-8}$	$2.3 \cdot 10^{-8}$	$1.6 \cdot 10^{-9}$
	Q3	$-7.9 \cdot 10^{-10}$	$6.8 \cdot 10^{-8}$	$2.9 \cdot 10^{-8}$	$3.1 \cdot 10^{-9}$

For each pair (α, γ), the quartiles (Q1, Q2, Q3, where Q2 is the median) of the maximum likelihood estimates of the fitness effect s and of the HGT rate h_β 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

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 1,000 samples, respectively. Based on our criterium for the log-likelihood described above, we find that the exponent combinations $(\alpha, \gamma) = (0, 1)$ (in all 1,000 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 γ suggest that if the assumptions of the model are correct, the transposition rate per IS*5* copy does not decrease with increasing IS*5* count per genome. The fitness exponent parameter α 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 draw conclusions about possible interactions between IS*5* 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 compute the total population size $Z = 1 + \frac{\hat{s}}{r} S_\alpha(\mathbf{q})$ in equilibrium, as well as the maximum likelihood estimate of the HGT rate, $\hat{h} = \frac{\hat{H}}{S_\beta(\mathbf{q})\, Z}$, which depends on our choice of the HGT exponent β. Table 2 shows for the four exponent pairs of $(\alpha, \gamma) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}$ that we found above the quartiles of the maximum likelihood estimates of s and of h for different 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 IS*5*.

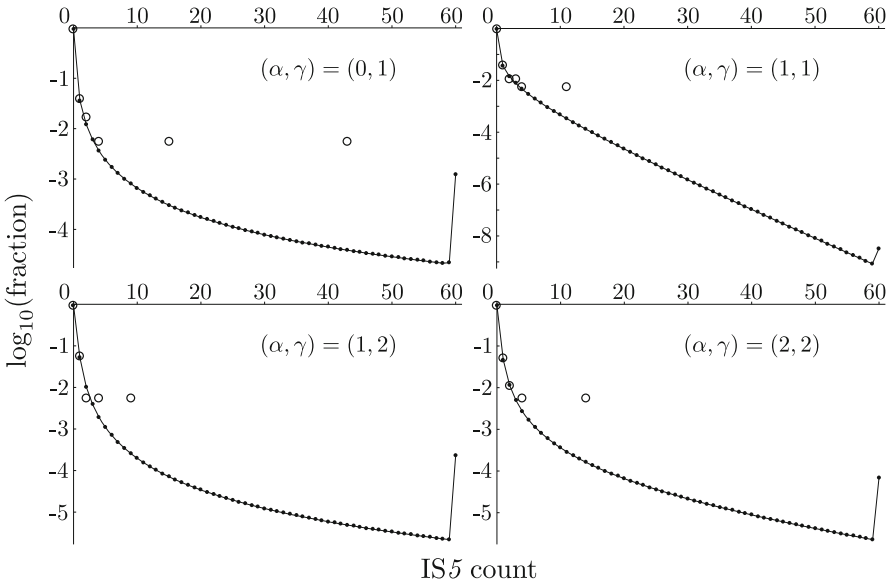

Fig. 3 Observed (*large circles*) and predicted (*small dots*, connected by a solid line) IS*5* count distributions for $(\alpha, \gamma) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}$. In the predicted IS count distribution, the IS count per genome has been limited to $l = 60$ copies of IS*5*. Note the logarithmic scale on the *vertical axis*. $\log_{10}(1/180) \approx -2.3$, i.e. each *large circle* at $\log_{10}$(fraction) $= -2.3$ represents only one genome

Table 2 shows that $\hat{s} < 0$ for IS*5*, i.e. that IS*5* 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, $\hat{h_\beta} \geq |\hat{s}|$. If, on the other hand, $\beta > \alpha$, then $\hat{h_\beta} < |\hat{s}|$ (not shown). The IS*5* infection, starting with cells carrying one copy of IS*5* only, will then die out, because HGT is not strong enough to overcome the negative fitness effect caused by even only one IS*5* copy per genome. Therefore, for an IS infection to spread, persist, and reach the observed IS*5* count distribution, the increase in the infectiousness of a cell with increasing IS count must be smaller than the simultaneous increase in the total fitness cost.

Figure 3 shows for each of the four exponent pairs $(\alpha, \gamma) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}$ an example of the predicted equilibrium distribution based on the maximum likelihood estimates $\hat{s}$ and $\hat{H}$, together with an observed IS*5* count distribution based on a sample that led to the best model fit with the chosen pair (α, γ) and the maximum likelihood estimates $\hat{s}$ and $\hat{H}$. The four pairs (α, γ) 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 $l = 60$ IS copies per genome. The bin with 60 copies per genome therefore represents all genomes with *at least* 60 copies in the computed distribution. (The highest IS*5* 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 $l = 60$ IS copies per genome. In a model with no upper bound for the IS count per genome,

the distribution would drop monotonously. We have confirmed this by using higher IS count limits l and by observing that the spike in the highest IS 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 IS*5* 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 or not at all on the cell genome's IS count. 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 $\frac{1}{2} \sum_{j=1}^{l} |Z_j(t)/Z_{\text{inf}}(t) - Z_j^\star/Z_{\text{inf}}^\star| = 0.1$ for the time t, where $Z_{\text{inf}} = \sum_{j=1}^{l} Z_j$ and an asterisk (*) indicates the final normalized population densities. In the appendix, we show for $(\alpha, \gamma) = (0, 1)$, $s = -1.1 \cdot 10^{-7}$ and $h_0 = 1.1 \cdot 10^{-7}$ the computed dynamics over time of a population of host cells infected with a fraction of 10^{-6} cells harboring one copy of an IS in their genome (see Fig. 5). We also demonstrate the effect of changing the transposition rate u, the IS excision rate e, the fitness effect s, and the HGT rate h_0 on the population dynamics and on the final IS count distribution (see Fig. 6).

We have computed the dynamics of the total host population size over time during an infection for each of the four exponent pairs (α, γ), using the same maximum likelihood estimates for s and 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 IS*5* is generally small, as our computations show.

3.3 The maximum likelihood estimates of the HGT rate and of the fitness effect are highly correlated

Using 1,000 random samples of 180 out of 525 proteobacterial genomes, one per genus, provides insights into the uncertainty about the real IS*5* count distribution, and the resulting uncertainty in determining exponent pairs (α, γ), fitness effect s, and HGT rate h. To also get information about the variation in the maximum likelihood

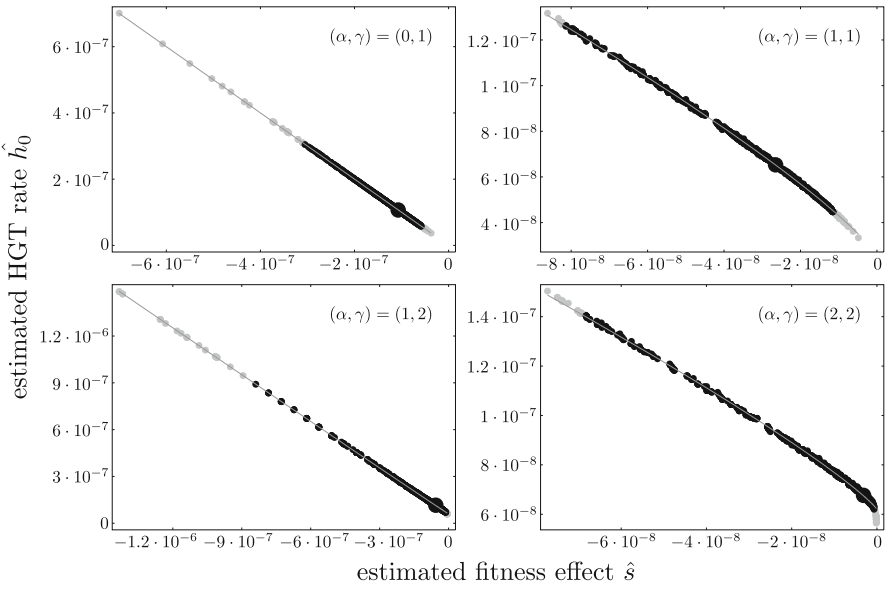

Fig. 4 Bootstrapped pairs of $(\hat{s}, \hat{h_0})$ for $(\alpha, \gamma) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}$, based on 1,000 resamplings of four computed IS count distributions. These four IS count distributions have been generated by using four estimate pairs $(\hat{s}, \hat{H})$, where each pair has been obtained from the IS*5* count distribution in a different random sample. *Black dots* lie inside and *gray dots* lie outside the 95% confidence interval of s. The original, estimated values of $(\hat{s}, \hat{h_0})$ 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 $\left(0.998, 1.000, 9.8 \cdot 10^{-12}\right)$ for $(\alpha, \gamma) = (0, 1)$, $\left(0.065, 0.819, 2.5 \cdot 10^{-8}\right)$ for $(\alpha, \gamma) = (1, 1)$, $\left(1.113, 1.008, 6.0 \cdot 10^{-8}\right)$ for $(\alpha, \gamma) = (1, 2)$, and $\left(0.114, 0.860, 6.2 \cdot 10^{-8}\right)$ for $(\alpha, \gamma) = (2, 2)$

estimates of s and h to be expected if our model were correct, and to gain some insight into the relationship between s and h, we use a bootstrap (Efron and Tibshirani 1994, p. 170). For each exponent pair $(\alpha, \beta) \in \{(0, 1), (1, 1), (1, 2), (2, 2)\}$, we choose the set of maximum likelihood estimates of s and H that led to the computed IS count distributions shown in Fig. 3. Based on each of the four computed IS count distributions, we then generated 1,000 artificial data sets and determined for each data set the maximum likelihood estimates of s and h_0. Figure 4 shows the values of $\hat{h_0}$ versus $\hat{s}$. We again show only the graphs for $\beta = 0$, as in the preceding subsection, and we 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.

The 1,000 bootstrapped pairs of $(\hat{s}, \hat{h_0})$ in Fig. 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, at least inside the 95% confidence interval. 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

$$h = \frac{ep_1 - sp_0 S_\alpha(\mathbf{p})}{p_0 Z S_\beta(\mathbf{p})} \approx -\frac{S_\alpha(\mathbf{p})}{S_\beta(\mathbf{p})} s. \tag{5}$$

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 an IS and in allowing an IS to persist in a host cell population.

4 Discussion

While an IS that provides a benefit to its host can rise to fixation through natural selection (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 infection. Specifically, we investigate whether a purely detrimental IS can persist and reach the observed IS*5* count distribution in proteobacteria, where IS*5* mainly occurs. We are also interested in the fitness effect s and in the HGT rate h needed to reach this IS count distribution. We find the maximum likelihood estimates of s and h_0 by analysing 525 fully sequenced genomes from 180 proteobacterial genera. We now discuss the main points of this study.

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

The L-shaped IS*5* count distribution in 525 sequenced proteobacterial genomes (and presumably also in natural host cell populations) suggests that IS*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*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 (Bichsel et al. 2010). 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). If so, the influence of the IS count on the HGT rate, and therefore on the infectiousness of a cell, is probably small or even absent.

4.2 The observed IS*5* count distribution suggests that the replicative transposition rate of IS*5* is not down-regulated

Our model shows best agreement with the IS*5* count distribution in 797 of 1,000 random samples, each containing 180 out of 525 proteobacterial genomes, if the

replicative transposition rate increases linearly with the IS*5* count per genome, i.e. if replicative transposition is not regulated and copies of IS*5* transpose independently. This is in agreement with results published by Sawyer et al. (1987). Using branching processes to model the count distribution of several ISs in the ECOR collection of 71 natural isolates of *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*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.

4.3 ISs might be effectively neutral to their hosts

Our model predicts a fitness effect in the range $\hat{s} \in \left[-10^{-7}, -10^{-9}\right]$ for IS*5* (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), IS*5* might therefore be effectively neutral or only slightly detrimental to its hosts. Hence, HGT is probably strong enough to enable IS*5* to persist and spread in a host cell population (see Table 1). At the same time, our model predicts an unrealistically long time for IS*5* to approximately reach the final prevalence of infected cells, while the predicted time to approximately reach the IS*5* count distribution in infected cells only is much shorter. It therefore seems that the time scale of the infection process may be much larger than the time scale of the process that leads to an equilibrium distribution in the population of infected cells. While the former time scale is determined by the antagonistic actions of HGT and the fitness cost of one IS copy, the latter time scale is determined mainly by replicative transposition and the fitness cost of varying numbers of IS copies. This observation of different time scales leads us to suggest that IS*5* 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.

4.4 Caveats

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 IS*5* count distribution in the sequenced genomes with the IS count distribution that our model of a single population predicts. However, we note that a very similar, L-shaped IS*5* 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 observation, together with the fact that the L-shaped IS count distribution can be observed in several 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 does also exist in other ISs and on the smallest taxonomic scale, that of a cell population.

Another objection to our approach might be that we use IS*5* count data from phylogenetically related genomes to conduct a maximum likelihood analysis which assumes independence between observations. The genomes in our data set are related and their IS*5* counts are therefore not strictly independent of each other. Nevertheless, we have reduced this dependence by choosing only one genome per genus for the likelihood analysis. Furthermore, we generated 1,000 sample data sets, each containing one genome per genus and repeated the maximum likelihood analysis for each of these data sets.

It might also be argued that the IS*5* count distribution in our data set is L-shaped because many ISs show certain DNA target specificities (Chandler and Mahillon 2002). IS*5* does in fact show some preference for the target sequence CTAG. However, because this nucleotide sequence is very short and therefore occurs frequently in host genomes, target specificity is probably not strong enough to limit the IS*5* count distribution noticeably in the IS count range on which we base our computations (0–60 copies of IS*5* per genome). This is supported by the observation that although most infected proteobacterial genomes have very low IS*5* counts, some genomes contain more than 40 copies of IS*5*. The same argumentation probably also holds for other ISs with some target specificities.

Acknowledgments 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, namely $(u, e) = \left(10^{-9}, 10^{-11}\right)$. Because the effect of the excision rate is small, and because the effect of an IS*5* 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 exponent pairs of (α, γ) 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 IS*5*.

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: IS*5* seems to be effectively neutral (even more so for this parameter combination of u and e), and HGT ist most

Table 3 For a replicative transposition rate $u = 10^{-9}$ and an excision rate $e = 10^{-11}$, the table shows the four most frequent exponent pairs $(\alpha, \gamma) \in \{0, 1, 2\}^2$ that lead to model fits of the IS*5* count distribution that are not significantly worse than the best fit

(α, γ)	Quart.	$\hat{s}$	$\hat{h_0}$	$\hat{h_1}$	$\hat{h_2}$
(0, 1)	Q1	$-1.6 \cdot 10^{-9}$	$1.1 \cdot 10^{-9}$	–	–
	Q2	$-1.3 \cdot 10^{-9}$	$1.3 \cdot 10^{-9}$	–	–
	Q3	$-1.1 \cdot 10^{-9}$	$1.6 \cdot 10^{-9}$	–	–
(1, 1)	Q1	$-2.8 \cdot 10^{-10}$	$3.8 \cdot 10^{-10}$	$7.3 \cdot 10^{-11}$	–
	Q2	$-1.8 \cdot 10^{-10}$	$5.5 \cdot 10^{-10}$	$1.8 \cdot 10^{-10}$	–
	Q3	$-7.3 \cdot 10^{-11}$	$6.7 \cdot 10^{-10}$	$2.8 \cdot 10^{-10}$	–
(1, 2)	Q1	$-2.9 \cdot 10^{-10}$	$7.8 \cdot 10^{-10}$	$1.9 \cdot 10^{-10}$	–
	Q2	$-2.3 \cdot 10^{-10}$	$8.2 \cdot 10^{-10}$	$2.3 \cdot 10^{-10}$	–
	Q3	$-1.9 \cdot 10^{-10}$	$8.8 \cdot 10^{-10}$	$2.9 \cdot 10^{-10}$	–
(2, 2)	Q1	$-3.1 \cdot 10^{-11}$	$6.4 \cdot 10^{-10}$	$1.8 \cdot 10^{-10}$	$8.0 \cdot 10^{-12}$
	Q2	$-1.6 \cdot 10^{-11}$	$6.6 \cdot 10^{-10}$	$2.4 \cdot 10^{-10}$	$1.6 \cdot 10^{-11}$
	Q3	$-7.9 \cdot 10^{-12}$	$6.8 \cdot 10^{-10}$	$2.9 \cdot 10^{-10}$	$3.1 \cdot 10^{-11}$

For each pair (α, γ), the quartiles (Q1, Q2, Q3, where Q2 is the median) of the maximum likelihood estimates of the fitness effect s and of the HGT rate h_β 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

probably strong enough to overcome the fitness cost caused by a copy of IS*5* in the host cell genome.

Population dynamics of an IS infection in dependence of the model parameter set

Figure 5 shows the computed population dynamics of a host cell population that has been infected with a fraction of 10^{-6} cells harboring one IS copy in their genomes. We chose $r = 1, u = 10^{-7}, e = 10^{-10}$, and the maximum likelihood estimates $\hat{s} = -1.1 \cdot 10^{-7}$ and $\hat{h_0} = 1.1 \cdot 10^{-7}$ for the exponent pair $(\alpha, \gamma) = (0, 1)$ to compute the infection dynamics based on the equation system 1.

Observe that the IS count distribution at 10^{13} generations in Fig. 5 is the same as the computed IS count equilibrium distribution in Fig. 2. As can be seen in Fig. 5, on the one hand, it takes a very long time to reach the population equilibrium of uninfected and infected cells. On the other hand, it takes a much shorter time to reach an equilibrium in the IS count distribution among infected cells only.

To illustrate the influence of different model parameters on the population dynamics and on the final IS count distribution, Fig. 6 shows the population dynamics if each of the parameters u, e, s, and h_0 has been separately set to one tenth of its original value as used in Fig. 5.

Compared with Figs. 5, 6 shows that reducing the transposition rate leads to a steeper final IS count distribution, with less cells harboring high numbers of IS copies in their genome (top left graph with $u = 10^{-8}$). Figure 6 also shows that reducing

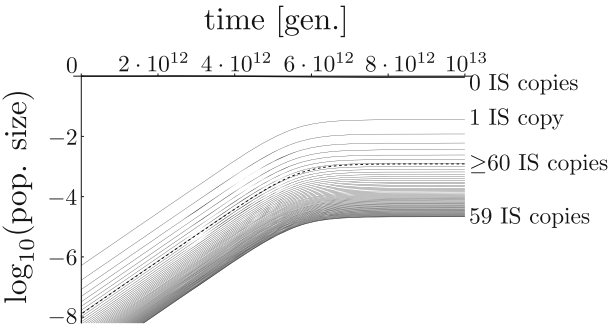

Fig. 5 Computed population dynamics of a host cell population infected with a fraction of 10^{-6} cells harboring one IS copy in their genomes. We chose $r = 1, u = 10^{-7}, e = 10^{-10}$, $(\alpha, \gamma, \beta) = (0, 1, 0)$, and the corresponding maximum likelihood estimates $\hat{s} = -1.1 \cdot 10^{-7}$ and $\hat{h_0} = 1.1 \cdot 10^{-7}$ as model parameters. The *curves* for cells harboring different numbers of IS copies in their genomes are indicated on the *right*. The *curve* for cells harboring 0 IS copies in their genomes is shown in *bold*, and the *curve* for cells harboring at least 60 IS copies in their genomes is shown as a *dashed line*. Time is measured in cell generations. Note the logarithmic scale on the *vertical axis*

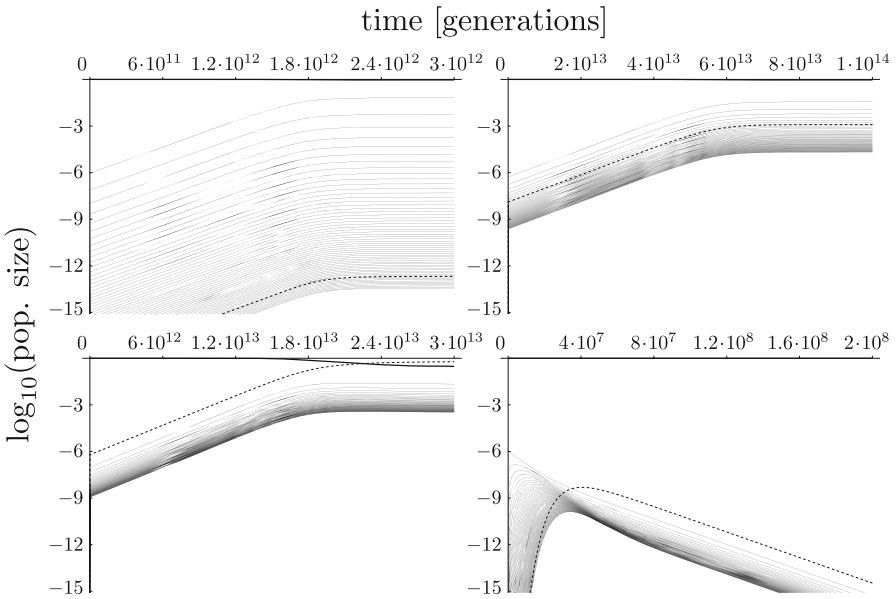

Fig. 6 Computed population dynamics of a host cell population infected with a fraction of 10^{-6} cells harboring one IS copy in their genomes, for different parameter sets. The original model parameters are the same as in Fig. 5: $r = 1, u = 10^{-7}, e = 10^{-10}$, $(\alpha, \gamma, \beta) = (0, 1, 0)$, $s = -1.1 \cdot 10^{-7}$, and $h_0 = 1.1 \cdot 10^{-7}$. For each of the four graphs, exactly one parameter has been changed compared to the original parameter set: $u = 10^{-8}$ (*top left*), $e = 10^{-11}$ (*top right*), $s = -1.1 \cdot 10^{-8}$ (*bottom left*), and $h_0 = 1.1 \cdot 10^{-8}$ (*bottom right*). In each graph, the curve for cells harboring 0 IS copies in their genomes is shown in *bold*, and the curve for cells harboring at least 60 IS copies in their genomes is shown as a *dashed line*. Time is measured in cell generations. Note the logarithmic scale on the *vertical axis*

the IS excision rate does not change the final IS count distribution noticeably, but it takes a longer time to reach this final distribution (top right graph with $e = 10^{-11}$). Reducing the fitness cost tenfold leads to a population dominated by infected cells with the highest IS count allowed in our model, noticeably reducing the normalized density of uninfected cells (bottom left graph with $s = -1.1 \cdot 10^{-8}$). Reducing the HGT rate below the fitness cost, in turn, does not allow the population of infected cells to persist (bottom right graph with $h_0 = 1.1 \cdot 10^{-8}$).

References

Berg DE (1989) Transposon Tn5. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D.C., pp 185–210

Bichsel M, Barbour AD, Wagner A (2010) The early phase of a bacterial insertion sequence infection. Theor Popul Biol 78:278–288

Blot M (1994) Transposable elements and adaptation of host bacteria. Genetica 93(1-3):5–12

Chandler M, Mahillon J (2002) Insertion sequences revisited. In: Craig NL, Craigie R, Gellert M, Lambowitz AM (eds) Mobile DNA II. American Society for Microbiology, Washington, D.C., pp 305–366

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215–219

Dahlberg C, Bergström M, Hermansson M (1998) In situ detection of high levels of horizontal plasmid transfer in marine bacterial communities. Appl Environ Microbiol 64(7):2670–2675

Dawkins R (1976) The selfish gene. Oxford University Press, Oxford

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601–603

Dröge M, Pühler A, Selbitschka W (1999) Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies. Biol Fertil Soils 29(3):221–245

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman & Hall/CRC, New York

Galas DJ, Chandler M (1989) Bacterial insertion sequences. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D.C., pp 109–162

Gibbons RJ, Kapsimalis B (1967) Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. J Bacteriol 93(1):510–512

Hall BG (1999) Transposable elements as activators of cryptic genes in *E. coli*. Genetica 107:181–187

Jiang SC, Paul JH (1998) Gene transfer by transduction in the marine environment. Appl Environ Microbiol 64(8):2780–2787

Kleckner N (1989) Transposon Tn10. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D.C., pp 227–268

Lan R, Reeves PR (2002) *Escherichia coli* in disguise: molecular origins of shigella. Microbes Infect 4(11):1125–1132

Lynch M (2007) The origins of genome architecture. Sinauer Associates, Inc, Sunderland

Madigan MT, Martinko JM, Dunlap PV, Clark DP (2009) Brock biology of microorganisms, 12th edn. Pearson Benjamin Cummings, San Francisco

Mahillon J, Siguier P, Chandler M (2009) IS Finder. http://www-is.biotoul.fr

NCBI (2011) National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov

Nelder JA, Mead R (1965) A simplex-method for function minimization. Comput J 7(4):308–313

Nuzhdin SV (1999) Sure facts, speculations, and open questions about the evolution of transposable element copy number. Genetica 107:129–137

Orgel LE, Crick FHC (1980) Selfish DNA: the ultimate parasite. Nature 284:604–607

Savageau MA (1983) *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. Am Naturalist 122(6):732–744

Sawyer SA, Dykhuizen DE, DuBose RF, Green L, Mutangadura-Mhlanga T, Wolczyk DF, Hartl DL (1987) Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics 115:51–63

Schneider D, Lenski RE (2004) Dynamics of insertion sequence elements during experimental evolution of bacteria. Res Microbiol 155:319–327

Seneta E (1981) Non-negative matrices and Markov chains. Springer, New York
Shapiro JA (1999) Transposable elements as the key to a 21st century view of evolution. Genetica 107: 171–179
So M, McCarthy BJ (1980) Nucleotide sequence of the bacterial transposon TN1681 encoding a heat-stable (ST) toxin and its identification in enterotoxigenic Escherichia coli strains. Proc Natl Acad Sci USA 77(7):4011–4015
Tavakoli NP, Derbyshire KM (2001) Tipping the balance between replicative and simple transposition. EMBO J 20(11):2923–2930
Top EM, Springael D (2003) The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. Curr Opin Biotechnol 14:262–269
Touchon M, Rocha EPC (2007) Causes of insertion sequences abundance in prokaryotic genomes. Mol Biol Evol 24(4):969–981
Wagner A (2006) Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. Mol Biol Evol 23(4):723–733
Wagner A, Lewis C, Bichsel M (2007) A survey of bacterial insertion sequences using IScan. Nucleic Acids Res 35(16):5284–5293
Williams HG, Day MJ, Fry JC, Stewart GJ (1996) Natural transformation in river epilithon. Appl Environ Microbiol 62(8):2994–2998
Wolfram S (2003) The mathematica book, 5th edn. Wolfram Media, Champaign