The early phase of a bacterial insertion sequence infection

Bichsel, M ; Barbour, A D ; Wagner, A

DOI: https://doi.org/10.1016/j.tpb.2010.08.003

The Early Phase of a Bacterial Insertion Sequence Infection

Manuel Bichsel*, Andrew D. Barbour†, Andreas Wagner‡

December 11, 2009

Abstract

Bacterial insertion sequences are the simplest form of autonomous mobile DNA. It is unknown whether they need to have beneficial effects to infect and persist in bacterial populations, or whether horizontal gene transfer suffices for their persistence. We address this question by using branching process models to investigate the critical, early phase of an insertion sequence infection. We find that the probability of a successful infection is low and depends linearly on the difference between the rate of horizontal gene transfer and the fitness cost of the insertion sequences. Our models show that the median time to extinction of an insertion sequence that dies out is very short, while the median time for a successful infection to reach a modest population size is very long. We conclude that horizontal gene transfer is strong enough to allow the persistence of insertion sequences, although infection is an erratic and slow process.

Keywords: Mobile DNA, Insertion sequence, Horizontal gene transfer, Branching process

1 Introduction

Ever since its discovery in the 1940s by Barbara McClintock [McClintock, 1950], mobile DNA has fascinated researchers. Why does it exist, and how does it persist? Some authors claim that mobile DNA ultimately needs to have beneficial effects on the host cell to be able to persist in the long term [Blot, 1994, Shapiro, 1999, Schneider and Lenski, 2004]. Other authors disagree and think that mobile DNA is selfish DNA, which merely persists by replicating inside a host cell's genome and by infecting new hosts through sexual reproduction or horizontal gene transfer [Dawkins, 1976, Doolittle and Sapienza, 1980, Orgel and Crick, 1980, Charlesworth et al., 1994, Nuzhdin, 1999]. While even purely detrimental mobile DNA can spread in a sexually reproducing eukaryote population [Charlesworth et al., 1994], the persistence of detrimental mobile DNA in an asexually reproducing, prokaryote population is more difficult to explain.

Besides raising theoretical issues, the existence and persistence of certain classes of mobile DNA is also of practical interest. Some prokaryotic transposons – mobile DNA elements that move inside their host genome through a cut-and-paste process – carry antibiotic resistance genes [Berg, 1989, Kleckner, 1989], genes encoding toxins [So and McCarthy, 1980], or genes with new metabolic functions [Top and Springael, 2003]. Thus, transposons on the one hand contribute to an important public health threat by spreading antibiotic resistance among pathogens. On the other hand, transposons are also very useful tools in genetic engineering.

*Department of Biochemistry, University of Zürich, 8057 Zürich, Switzerland
†Institute of Mathematics, University of Zürich, 8057 Zürich, Switzerland
‡Department of Biochemistry, University of Zürich, 8057 Zürich, Switzerland

Prokaryotic transposons consist of two groups: simple and composite transposons. Simple transposons encode the proteins needed for their mobility themselves. Composite transposons contain two flanking insertion sequences, another class of mobile DNA. The insertion sequences encode the protein needed for the composite transposon's mobility.

Bacterial insertion sequences (ISs) are short DNA segments with a length of between 700 and 2700 bp [Chandler and Mahillon, 2002]. An IS usually codes for only one protein, transposase, which excises it from its current position in the genome and inserts it at a new position, a process called conservative transposition. Occasionally, instead of being cut-and-pasted, an IS is copy-and-pasted through replicative transposition. Replicative transposition increases the IS count per genome; however, ISs are sometimes also excised, thus decreasing the IS count. ISs are probably the simplest form of autonomous mobile DNA, encoding for just enough functionality to move and spread on their own inside a host genome. Currently, all ISs have been classified into 20 families, based on differences in their internal organization (open reading frames), in their transposases, in the nucleotide sequence at their ends, and in the nucleotide sequences they leave behind in the genome after being excised [Chandler and Mahillon, 2002, Mahillon et al., 2009]. Individual ISs are named ISn, where n is an integer (e.g. IS*1*, IS*2* and IS*630*).

ISs pose a threat to host cells for at least two reasons. First, ISs can disable genes by inserting themselves into them. Second, if more than one IS is present in a genome, ISs can lead to the rearrangement of the whole host genome through homologous recombination [Galas and Chandler, 1989, Kleckner, 1989, Schneider and Lenski, 2004]. Therefore, although ISs can occasionally cause beneficial mutations [Hall, 1999, Schneider and Lenski, 2004], their general effect on the host cell is probably detrimental, especially if the IS count per genome is high.

Why then do ISs persist? When an IS first enters an uninfected host cell population, it occurs in only one or a few genomes of the population. It can then spread by horizontal gene transfer (HGT) to the genomes of other cells. This early phase of an IS invasion is crucial for its long-term fate and has parallels in the fate of a rare, slightly detrimental allele in a large population [Ohta, 1974]. HGT is a necessary condition for the persistence of a detrimental IS. But is HGT enough to allow IS persistence, or are (albeit rare) beneficial effects of ISs needed? We address this question by modeling the early phase of an invasion of a slightly detrimental IS into an uninfected bacterial host cell population as a branching process. Specifically, we first use our branching process models to compute the survival probability of an IS infection, its time to extinction if it becomes extinct, and the time to reach a given population size threshold if the infection persists. Last, we use our multi-type branching process model to derive the distribution of the IS count per infected cell genome, and we compare this distribution with the real IS count distribution in 728 fully sequenced bacterial genomes.

2 Models

In our models, we assume a large bacterial host cell population living at carrying capacity. Into this population, we introduce one cell infected with a single IS. We use a continuous-time, multi-type Markov branching process model to compute the IS survival probability, the time needed to reach a given population size threshold if the IS persists, and the IS count distribution [Haccou et al., 2005, Athreya and Ney, 1972]. We use a related birth-and-death process model to analyse the time to extinction if the IS becomes extinct. Being stochastic processes, branching processes are particularly well-suited to model the early phase of an IS infection, given that the number of infected cells is still low and prone to strong random fluctuation. The use of branching process models in population genetics dates back to Fisher [Fisher, 1922] and Haldane [Haldane, 1927]. For introductions to branching processes and their use in biol-

ogy, see [Athreya and Ney, 1972, Sewastjanow, 1975, Jagers, 1975, Kimmel and Axelrod, 2002, Haccou et al., 2005].

As we only model the early phase of an IS infection, we assume that the number of infected cells is always several orders of magnitude lower than the number of uninfected cells. We furthermore assume the cells to live in a well-mixed bulk environment, e.g. in seawater. In such an environment, each infected cell is surrounded by uninfected cells only and not influenced by any other infected cells, i.e. there is no HGT between infected cells.

We do not allow for immigration or emigration of cells, and as the host cell population lives at carrying capacity, the cell division rate b equals the base death rate d. For convenience, we choose $b = d = 1$ per cell generation. This choice of the cell division rate leads to the generation time being one time unit. As ISs are relatively short compared to their host genome (2.7 kbp at the most, versus e.g. around 4500 to 5500 kbp for the *E. coli* genome [Bergthorsson and Ochman, 1998]), we neglect the small additional cost needed in replicating ISs during cell division and assume the same birth rate b for infected cells as for uninfected cells. Empirical data suggest a death rate of infected cells with at most a linear dependence on the IS count per genome [Sawyer et al., 1987]. We assume a linearly increasing death rate of the form $d + js$ for infected cells, where j is the IS count per genome, and $s \ll d$ is the fitness cost per IS.

We allow for five event types that change the total IS count in the population: division of an infected cell, death of an infected cell, replicative transposition of an IS, excision of an IS, and HGT. In HGT, an IS is copied from an infected cell to an uninfected cell.

2.1 Multi-type model

Our multi-type model is inspired by and similar to the models used by Moody [Moody, 1988] and by Basten and Moody [Basten and Moody, 1991], but our model differs in the effect of a cell's IS count on the cell's fitness, and, more importantly, instead of assuming a fixed bacterial generation time, we assume a continuous, exponentially distributed generation time. Although not strictly correct [Powell, 1955], an exponentially distributed generation time has been chosen to simplify calculations, because the branching process is then also a Markov process. In any case, our results will still be qualitatively correct if a better suited non-exponentially distributed generation time is assumed.

Some ISs down-regulate their transposition rate with increasing IS count per genome [Sawyer et al., 1987, Chandler and Mahillon, 2002]. An example is IS*10*, where the IS produces both a locally operating transposase and a globally operating negative regulator of transposase gene expression, so that with increasing IS count the transposase density at an IS site stays constant, while the density of the negative regulator increases. We include this effect in our model and assume the replicative transposition rate u per infected cell and per generation to be constant and independent of the cell genome's IS count (but see subsection 4.5 for a discussion of the effects of a nonconstant transposition rate). Furthermore, we assume excision events to be independent of each other. In our multi-type model, we therefore adopt a rate je of IS excision events per infected cell and generation, proportional to the genome's IS count j and the excision rate e per IS, where $e < u$ [Egner and Berg, 1981]. It is not known whether the IS count of a cell's genome influences the cell's HGT rate. But it is known that HGT is tightly regulated and depends on many internal and external factors [Dröge et al., 1999], of which the IS count of the donor cell is probably only a minor one. For simplicity, we assume a constant rate h of HGT per infected cell and per generation, independent of the cell genome's IS count (see subsection 4.5 for a discussion of the effects of a nonconstant HGT rate).

To avoid having to deal with an infinite-dimensional system, we assume an upper limit of $l = 50$ ISs per genome, except where noted otherwise. This is not a serious restriction, because only a very small proportion of infected cells in the wild has such a high IS count, and most infected

cells harbor only a few ISs in their genome, as has already been seen before [Sawyer et al., 1987, Wagner, 2006, Touchon and Rocha, 2007], and as we also show in subsection 3.4.

Figure 1 shows the structure of the multi-type model, as defined by our assumptions.

[insert figure 1 here]

A cell genome's IS count k, $k \in \{1, \ldots, l\}$, determines the cell's event rate a_k, i.e. the rate at which either a cell death, a cell birth, a replicative transposition event, an excision event, or an HGT event happen in a cell harboring k ISs:

$$\begin{aligned} a_1 &= b + d + s + u + e + h \\ a_j &= b + d + js + u + je + h \quad (1 < j < l) \\ a_l &= b + d + ls + le + h, \end{aligned}$$

where b and d are the birth and base death rates, respectively, s is the fitness cost per IS copy, u is the replicative transposition rate, e is the IS excision rate, and h is the rate of HGT.

The waiting time to the cell's next event is assumed to have an exponential distribution with mean $1/a_k$, and at the time of an event, the probabilities p_k of the five different event types are given by

IS count	cell div.	cell death	transp.	excision	HGT
1	$\frac{b+h}{a_1}$	$\frac{d+s+e}{a_1}$	$\frac{u}{a_1}$	0	0
$1 < j < l$	$\frac{b}{a_j}$	$\frac{d+js}{a_j}$	$\frac{u}{a_j}$	$\frac{je}{a_j}$	$\frac{h}{a_j}$
l	$\frac{b}{a_l}$	$\frac{d+ls}{a_l}$	0	$\frac{le}{a_l}$	$\frac{h}{a_l}$.

For a cell infected with one IS, excision is counted as cell death (uninfected cells are not included in the model), and HGT is counted as cell division.

The event probabilities p_k are then used to define the vector-valued probability generating function $\mathbf{g}(\mathbf{z}) = \sum_{\mathbf{j}} \mathbf{p}(\mathbf{j})\mathbf{z}^{\mathbf{j}}$, $\mathbf{z} = (z_1, \ldots, z_l)$ (see A). From the probability generating function, we derive the infinitesimal generating functions $\tilde{g}_k(\mathbf{z}) = a_k(g_k(\mathbf{z}) - z_k)$, and the infinitesimal generator A, which is defined as $A = (a_{ij}) = a_i b_{ij}$, where $b_{ij} = \left.\frac{\partial g_i(\mathbf{z})}{\partial z_j}\right|_{\mathbf{z}=\mathbf{1}} - \delta_{ij}$ [Athreya and Ney, 1972, p. 183 and 200], also shown in A. The eigenvalue λ_0 of A with the largest real part is itself real. If λ_0 is negative, 0, or positive, the branching process is called subcritical, critical, or supercritical, respectively. If the branching process is subcritical or critical, it will become extinct with certainty; if the branching process is supercritical, it has a positive probability smaller than one of survival.

If the branching process is supercritical, there exist positive right and left eigenvectors $\mathbf{u} = (u_1, \ldots, u_l)$ and $\mathbf{v} = (v_1, \ldots, v_l)$ of the infinitesimal generator A, which can be scaled so that $\sum_{k=1}^{l} u_k = 1$ and $\sum_{k=1}^{l} u_k v_k = 1$. In the following, it will always be assumed that this scaling has been done for $\mathbf{u}$ and $\mathbf{v}$.

2.2 Single-type model

For the birth-and-death process model, we simplify the multi-type model by assuming that transposition and excision can be neglected, so that there is only one type of infected cell, bearing exactly one IS. The process state of the birth-and-death process model corresponds to the number of infected cells, and process state 0 is considered to be absorbing, meaning that the population of infected cells has become extinct. The birth and death rates per infected cell are $b + h$ and $d + s$, respectively, where again b and d are the birth and base death rates of a cell, h is the HGT rate, and s is the fitness cost of an IS.

Feller was the first to investigate this birth-and-death process [Feller, 1939]. Kendall derived the probability $P_n(t)$ of the process being in state n at time t [Kendall, 1948]. In our case, this probability is

$$P_n(t) = \begin{cases} \xi_t & \text{if } n = 0 \\ (1-\xi_t)(1-\eta_t)\,\eta_t^{n-1} & \text{if } n > 0 \end{cases} \quad (1)$$

where

$$\xi_t = \frac{(d+s)\left(1-e^{-(b+h-(d+s))t}\right)}{b+h-(d+s)\,e^{-(b+h-(d+s))t}} \text{ and } \eta_t = \frac{(b+h)\left(1-e^{-(b+h-(d+s))t}\right)}{b+h-(d+s)\,e^{-(b+h-(d+s))t}}.$$

At all times t, the state of the birth-and-death process is therefore zero (i.e. the process has become extinct) with probability ξ_t, and otherwise the state has a geometric distribution with parameter η_t.

2.3 Model parameters

We now turn to the parameters that we are using to analyze the models. Reliable rates for replicative transposition, IS excision and HGT are difficult to establish. However, in some cases at least their order of magnitude is known or can be estimated. Conservative transposition occurs with a rate of around 10^{-7} to 10^{-4} events per cell and host cell generation [Chandler and Mahillon, 2002, Kleckner, 1989]. We assume the replicative transposition rate to be a few orders of magnitude smaller [Tavakoli and Derbyshire, 2001]. IS excision rates are lower than replicative transposition rates [Egner and Berg, 1981]. For example, IS*10* is excised from the genome at a rate of around 10^{-10} per cell and host cell generation, whereas its conservative transposition rate is 10^{-4} per cell and host cell generation [Kleckner, 1989]. Similarly, transposon Tn*5*, a mobile DNA sequence flanked by two copies of IS*50*, has an excision rate of 10^{-6} to 10^{-5} and a conservative transposition rate of 10^{-3} to 10^{-2} [Berg, 1977]. HGT rates vary widely and depend on many environmental factors. For viral transduction in marine bacteria, rates of between $1.6 \cdot 10^{-8}$ and $3.7 \cdot 10^{-8}$ transductants per colony-forming unit have been reported [Jiang and Paul, 1998]. For the conjugational transfer of plasmids in diverse seawater bacteria, $2.3 \cdot 10^{-6}$ to $5.6 \cdot 10^{-5}$ transconjugants per recipient cell have been found after 3 days of incubation [Dahlberg et al., 1998]. For transformation involving epilithic bacteria from a river, *in situ* rates of $2.2 \cdot 10^{-6}$ to $1.0 \cdot 10^{-3}$ events per recipient cell have been reported per 24 hours incubation time [Williams et al., 1996]. Note that in this case, the transformation occurred in cells that were fixed on a surface, i.e. not in a well-mixed environment as we assume in our models. No information is available about the fitness cost caused by ISs. In our models, we therefore vary this cost over a broad range of values.

Table 1 shows a summary of reported rates and rates used in our models.

[insert table 1 here]

2.4 Software

We use Mathematica version 7.0 to carry out the numerical and analytical model computations. With the exception of figure 7, we also use Mathematica to generate the figures in the results section. Figure 7 has been generated by first counting ISs in fully sequenced bacterial genomes using IScan [Wagner et al., 2007], and then computing the IS count distribution using R, version 2.6.2 [R Development Core Team, 2008].

3 Results

3.1 The survival probability of an IS infection is small

The survival probability p_{surv} of an IS infection starting with one cell that is infected with a single IS is given by $p_{\text{surv}} = 1 - p_{\text{ext}}$, where p_{ext} is the infection's extinction probability. The extinction probability of an IS infection starting with one cell that is infected with k ISs is the k-th component of the smallest root $\mathbf{q} = (q_1, \ldots, q_l)$ of the infinitesimal generating function $\tilde{\mathbf{g}}(\mathbf{z})$ in the interval $[\mathbf{0}, \mathbf{1}]$ [Athreya and Ney, 1972, p. 205]. The survival probability of an infection starting with one cell that contains one IS in its genome can therefore be computed as $p_{\text{surv}} = 1 - q_1$.

Figure 2 shows the survival probability as a function of the relative difference between the HGT rate and the fitness cost of an IS, based on a numerical computation of q_1 for different parameter combinations.

[insert figure 2 here]

Figure 2 shows that $p_{\text{surv}} \approx h - s$, i.e. that the survival probability of an IS infection starting with one cell that is infected with one IS is approximately equal to the difference between the HGT rate and the fitness cost, at least if the replicative transposition rate u is smaller than the fitness cost s per IS. Only if $u > s$ does the infection's survival probability drop well below $h - s$ for low HGT rates h. The comparatively small excision rate does not have a significant effect on the infection's survival probability.

This result can be interpreted as follows: an IS infection can only persist if HGT is strong enough to overcome the mean fitness cost induced by ISs in infected cells (cf. figure 1). For replicative transposition rates that are lower than the fitness cost per IS, most cells will have only one IS. In that case, the survival probability of an infection will linearly depend on the difference $h - s$ between the HGT rate and the fitness cost induced by one IS. If, on the other hand, the replicative transposition rate is much larger than the fitness cost per IS, the population of infected cells includes many cells with higher IS counts, thus increasing the mean fitness cost per infected cell. This leads to a survival probability lower than $h - s$.

The negative effect that a high replicative transposition rate has on the survival probability of an IS infection can also be demonstrated by computing the HGT rate h_{crit} at which the multi-type branching process is critical and will only just become extinct with certainty. h_{crit} can be computed by observing that λ_0, the eigenvalue with the largest real part of the infinitesimal generator A (see A), must then be 0. Therefore, the constant term in the characteristic polynomial of A, which equals the determinant of A, must vanish. As h occurs only in the first column of A, the constant term linearly depends on h, and looking for its root, we find h_{crit}. Figure 3 shows h_{crit} as a function of s.

[insert figure 3 here]

Figure 3 shows that for a fitness cost much larger than the replicative transposition rate u (infected cells then carry only one IS), the critical HGT rate is equal to the fitness cost. Figure 3 also shows that for a fitness cost coming near or falling below the replicative transposition rate (infected cells then carry on average more than one IS), the critical HGT rate is higher than the fitness cost per IS, because HGT has to compensate for a larger total fitness cost caused by a higher IS count per cell.

We will see later that the IS count distribution in infected cells is indeed strongly L-shaped, i.e. most infected cells contain only one or at most a few ISs in their genome (see subsection 3.4). We can therefore use the birth-and-death process model as an approximation to our multi-type

branching process model. In this single-type model, we can analytically confirm that $p_{\text{surv}} \approx h-s$ for small values of h and s. To do this, we observe that our birth-and-death process only survives if it does not get absorbed in state 0. Using (1), we therefore get

$$p_{\text{surv}} = 1 - \lim_{t\to\infty} P_0(t) = 1 - \frac{d+s}{b+h}.$$

Remembering that $b = d = 1$ gen^{-1}, and linearizing around $h = s = 0$ gen^{-1} then gives

$$p_{\text{surv}} \approx h - s.$$

Haldane [Haldane, 1927], following an idea of Fisher [Fisher, 1922], showed that a dominant mutant gene with a small selective advantage s, so that the expected number of offspring is $1 + s$, has a probability of about $2s$ to persist in a random mating population. Observe that in our case, the selective advantage of a cell that harbors an IS is $(h - s)/2$, as the cell's expected number of offspring is $2 \cdot (b + h)/(b + d + h + s) \approx 1 + (h - s)/2$ for $b = d = 1$ gen^{-1} and small h and s.

3.2 The time to extinction of an IS infection is short

According to the last section, the vast majority of IS infections die out. Again considering that IS infections are dominated by cells with only a few ISs (see subsection 3.4), we use the single-type birth-and-death process model to compute the time to extinction of an IS infection that becomes extinct. We start with one infected cell in an uninfected host cell population. We then use the process state probability given in (1), observing that the probability of the birth-and-death process ever becoming extinct is given by $\lim_{t\to\infty} P_0(t)$. Therefore, using our assumption that $b = d = 1$ gen^{-1}, the cumulative distribution function F of the time to extinction T_0, conditioned on the branching process becoming extinct, is

$$F(t) = P(T_0 \le t | T_0 < \infty) = \frac{P_0(t)}{\lim_{t\to\infty} P_0(t)}.$$

As we have shown earlier, only in the case $h > s$ is there a positive probability of the birth-and-death process not becoming extinct. The distribution function is then

$$F(t) = \frac{(1+h)\left(1 - e^{-(h-s)t}\right)}{1 + h - (1+s)\, e^{-(h-s)t}},$$

and the corresponding probability density function of the time to extinction, conditioned on the branching process becoming extinct, is

$$f(t) = \frac{dF(t)}{dt} = \frac{(1+h)(h-s)^2\, e^{-(h-s)t}}{\left(1 + h - (1+s)\, e^{-(h-s)t}\right)^2}.$$

Figure 4 shows the density of T_0 for different parameter combinations of the fitness cost s and the HGT rate h, where always $h > s$.

[insert figure 4 here]

Figure 4 shows that first, the time to extinction is not strongly influenced by the fitness cost of an IS and by the HGT rate, and second, the distribution of the time to extinction is very skewed. Because of the latter observation, the median $T_{0,\text{med}}$ of the time to extinction is more useful to report than the mean. We use the distribution F of the time to extinction to obtain

the median time to extinction. To this end, we first transform F algebraically and then linearize the transformed expression around $h = s = 0$ gen^{-1}:

$$F(t) = \frac{1}{1 + \frac{\frac{h-s}{1+h}\, e^{-(h-s)t}}{1-e^{-(h-s)t}}} \approx \frac{t}{t+1} + \frac{1}{2}\left(\frac{t}{t+1}\right)^2 \left(\frac{t+2}{t}h - s\right)$$

Solving the equation $F(t) = 1/2$ for t and then again linearizing around $h = s = 0$ gen^{-1} gives the median time

$$T_{0,\text{med}} \approx \frac{\sqrt{1+h+h^2-s} - h}{1+h-s} \approx 1 - \frac{3h-s}{2}$$

The median time to extinction of an IS infection that becomes extinct therefore almost linearly depends on $3h - s$, but is dominated by the comparatively large constant 1. In this short time, replicative transposition and excision cannot take effect, which adds justification to our use of the birth-and-death process model.

3.3 The time an IS infection needs to attain a modest size threshold is long

Only a small fraction of IS infections survives. In a branching process, the surviving populations go into exponential growth after having lingered at lower population sizes during a random time period [Haccou et al., 2005, p. 158], [Athreya and Ney, 1972, p. 206], where they have been under strong threat of extinction. We first use our multi-type branching process model to numerically compute the time needed by a surviving population of infected cells to reach a given population size threshold. We then use our single-type birth-and-death process model to analytically confirm our numerical results from the multi-type model.

In a supercritical, irreducible, multi-type branching process with finite second moment as described by our multi-type model, the following holds [Sewastjanow, 1975, pp. 257–258]:

1. The random variable $W_k^m(t) := \frac{Z_k^m(t)}{v_k e^{\lambda_0 t}} \overset{t\to\infty}{\longrightarrow} W^m$ for any $m \in \{1, \ldots, l\}$ and $k \in \{1, \ldots, l\}$, where $Z_k^m(t)$ is the number of cells of type k at time t, starting with one cell of type m at time $t = 0$, and where v_k is the k-th component of the scaled left eigenvector $\mathbf{v}$ to the eigenvalue λ_0 of the infinitesimal generator A defined in A.

2. The characteristic function $\varphi^m(x) = \mathbb{E}\left(e^{iW^m x}\right)$ of W^m, where $i = \sqrt{-1}$, obeys the system of ordinary differential equations $\frac{d\varphi^m(x)}{dx} = \frac{\tilde{g}^m(\varphi^1(x),\ldots,\varphi^l(x))}{\lambda_0 x}$, with $\varphi^m(0) = 1$ for $m \in \{1, \ldots, l\}$, where $\tilde{g}^m$ is the infinitesimal generating function.

The ordinary differential equation system can be numerically solved for the characteristic functions $\varphi^m(x)$, $m \in \{1, \ldots, l\}$ (see B for details of the system). By the Fourier inversion theorem, the probability density f^1 of the random variable W^1 can be reconstructed from its characteristic function φ^1 as $f^1(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi^1(x)dx$. From W^1, in turn, the number $Z_k^1(t)$ of infected cells with k ISs at time t (large enough) in a population that has been infected with one cell containing one IS in its genome can be derived as $Z_k^1(t) \approx v_k e^{\lambda_0 t} W^1$. The total size of the population of infected cells is then $Z(t) := \sum_{k=1}^{l} Z_k^1(t) \approx e^{\lambda_0 t} W^1 \sum_{k=1}^{l} v_k$. Therefore, the time T_N to the threshold N is

$$T_N = \frac{1}{\lambda_0}\left[\ln(N) - \ln(W^1) - \ln\left(\sum_{k=1}^{l} v_k\right)\right].$$

We again use the median to characterize the time to threshold and get

$$T_{N,\text{med}} = \frac{1}{\lambda_0}\left[\ln(N) - \ln(W^1_{\text{med}}) - \ln\left(\sum_{k=1}^{l} v_k\right)\right],$$

where W^1_{med} is the median of the random variable W^1, which can be computed using the density f^1 of W^1. Figure 5 shows the median time to a threshold of 10^8 infected cells versus the difference between the HGT rate and the fitness cost, for different fitness costs s, replicative transposition rates u and excision rates e. $N = 10^8$ is still a comparatively small threshold in a population of bacterial cells. In a bulk environment like seawater, for example, $8.4 \cdot 10^8$ to $2.5 \cdot 10^{10}$ bacterial cells per liter have been counted [Thompson et al., 2004]. And still, the threshold is large enough to guarantee a negligible extinction probability once it has been attained by the population of infected cells. Because of the computational complexity involved in calculating the time to threshold, the maximal number of ISs per cell genome had to be reduced from $l = 50$ to $l = 5$. This is not a strong limitation, since the population of infected cells is dominated by cells that harbor only one or very few ISs.

[insert figure 5 here]

Figure 5 shows that the median time to threshold is approximately inversely proportional to $h-s$ for large thresholds N, e.g. $T_{N,\text{med}} = 55.5 \cdot (h-s)^{-0.82}$ for $s = 10^{-8}$ gen.$^{-1}$ and $N = 10^8$. We have confirmed that for larger thresholds the approximation to inverse proportionality becomes even better, e.g. $T_{N,\text{med}} = 35.5 \cdot (h-s)^{-0.93}$ for $s = 10^{-8}$ gen.$^{-1}$ and $N = 10^{12}$ (graph not shown). This is because first, for large thresholds N, the population dynamics of the supercritical branching process is dominated by the exponential growth phase; second, the time spent in the exponential growth phase is inversely proportional to the growth rate, which is identical to the eigenvalue λ_0 of the infinitesimal generator A; third, at least for h much larger than s if $u \geq s$, λ_0 is approximately equal to the difference $h - s$ between the HGT rate and the fitness cost (see figure 6).

[insert figure 6 here]

Figure 6 shows that if $s > u$ or $h > u$, $\lambda_0 \approx h - s$. Using a linear regression on the data shown in figure 6 that is restricted to $s = 10^{-8}$ gen.$^{-1}$ shows that $\lambda_0 \approx 1.00060 \cdot (h-s)^{1.00005}$. But λ_0 is smaller than $h - s$ (and can even drop below zero) if $u > s$ and $u \geq h$, because the population of infected cells is then no longer dominated by cells with only one IS, and HGT cannot replace fast enough the cells dying due to the increased total fitness cost per cell.

Because the population of infected cells is dominated by cells with only one IS, the single-type model is a good approximation to the multi-type model. We now use the birth-and-death process model to analytically show that the median time to threshold is in fact approximately inversely proportional to the difference between the HGT rate and the fitness cost. Let again $Z(t)$ be the size of the population of infected cells at time t. Then $Z(t)/e^{(b+h-(d+s))t}$ is a nonnegative Martingale, and thus $\lim_{t\to\infty} Z(t)/e^{(b+h-(d+s))t} = W$ almost surely exists [Athreya and Ney, 1972, p. 111]. W is a random variable that is zero with probability $P(W = 0) = \frac{d+s}{b+h}$ and otherwise has an exponential distribution with rate parameter $\frac{b+h-(d+s)}{b+h}$ [Harris, 1951, p. 319].

From $\lim_{t\to\infty} Z(t)/e^{(b+h-(d+s))t} = W$, we get $\ln(Z(t)) - (b + h - (d + s))t \approx \ln(W)$ if t is large. Therefore, the time T_N to reach the threshold N, on the condition that it is reached, is $T_N \approx \frac{1}{b+h-(d+s)}[\ln(N) - \ln(W)]$. Using this approximation, we get for T_N the distribution

function

$$\begin{aligned} P(T_N \le t) &= P\left(\frac{1}{b+h-(d+s)}(\ln(N)-\ln(W)) \le t\right) \\ &= 1 - P\left(W < N\, e^{-(b+h-(d+s))\,t}\right) \\ &= 1 - \int_0^{N\, e^{-(b+h-(d+s))\,t}} \frac{b+h-(d+s)}{b+h}\, e^{-\frac{b+h-(d+s)}{b+h}x}\, dx \\ &= \exp\left\{-\exp\left[-\left(x - \frac{\ln\left(N\frac{b+h-(d+s)}{b+h}\right)}{b+h-(d+s)}\right) \Big/ \frac{1}{b+h-(d+s)}\right]\right\}. \end{aligned}$$

This means that T_N has a Gumbel distribution, $P(T_N \le t) = \exp\left(-e^{-(x-a)/b}\right)$ with parameters $a = \frac{1}{b+h-(d+s)} \ln\left(N\frac{b+h-(d+s)}{b+h}\right)$ and $b = \frac{1}{b+h-(d+s)}$, see [Johnson et al., 1995, p. 2], and therefore the median time to threshold is

$$\begin{aligned} T_{N,\text{med}} &= \frac{1}{b+h-(d+s)} \ln\left(N\frac{b+h-(d+s)}{b+h}\right) - \frac{1}{b+h-(d+s)} \ln(\ln(2)) \\ &= \frac{1}{b+h-(d+s)} \left[\ln\left(N\frac{b+h-(d+s)}{b+h}\right) - \ln(\ln(2))\right] \\ &\approx \frac{1}{h-s} \ln(N) \quad \text{if } N \text{ big and } b = d. \end{aligned}$$

This result shows that for large population size thresholds, the median time to threshold is approximately inversely proportional to the difference $h - s$ between the HGT rate and the fitness cost, and that the proportionality constant is the natural logarithm of the threshold size N.

3.4 The IS count distribution is biased towards low IS counts

The IS count distribution is the link between our model and real data. We demonstrate that our multi-type branching process model can adequately reproduce the real IS count distribution. Figure 7 shows the IS count distribution of the six most abundant ISs IS*1*A, IS*2*, IS*4*, IS*5*, IS*110* and IS*630*, which occur in at least 20 of the 728 bacterial genomes that have been fully sequenced as of June 2009. We obtained the necessary genome sequences from the National Center for Biotechnology Information, NCBI [NCBI, 2009], and we obtained the reference sequences of the ISs from the IS Finder database [Mahillon et al., 2009]. We used our previously published software IScan to identify and count ISs in the genomes, analogous to our earlier work [Wagner et al., 2007], but for a larger number of genomes.

[insert figure 7 here]

Figure 7 shows that for each of the six most abundant ISs we examined, on average only 31 out of 728 sequenced bacterial genomes contain a minimum of one copy. The IS count distribution is L-shaped: most genomes contain none of these six ISs, a small number of genomes have up to a dozen copies of these ISs, and only a few genomes contain more than a dozen copies, although there are a few genomes containing many ISs. Among the six ISs we examined, only IS*1*A and IS*5* have more than 50 copies in some bacterial genomes: the seven sequenced *Shigella* genomes contain between 105 and 228 copies of IS*1*A, and *Xanthomonas oryzae* contains 53 copies of IS*5*. Of the 14 other, less abundant ISs we examined, only IS*481* and IS*982* have more than 50 copies in a genome: *Bordetella pertussis* contains 233 copies of IS*481* (all other genomes contain

at most 11 copies of IS*481*), and *Lactotcoccus lactis cremoris* contains 56 copies of IS*982* (all other genomes contain at most 3 copies of IS*982*).

We do not distinguish between different prokaryotic species in the data of figure 7, because, especially for prokaryotes, HGT occurs across species boundaries [Gogarten and Townsend, 2005, Sørensen et al., 2005]. It is known that many ISs show DNA target specificities of varying degrees [Chandler and Mahillon, 2002]. For example, while IS*1* just prefers AT-rich regions, IS*4* is known to insert into DNA sequences of the form AAA–N_{15-20}–TTT [Zerbib et al., 1985, Mayaux et al., 1984]. In practice, target specificity is probably not strong enough to be a limiting factor in the IS count distribution.

We now derive the model's IS count distribution by pointing out that for our multi-type branching process, the limit $\lim_{t\to\infty} \frac{\mathbf{Z}(t)}{e^{\lambda_0 t}} = W\mathbf{v}$ almost surely exists, where $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_l(t))$ is the vector of population sizes of infected cells with IS count $k \in \{1, \ldots, l\}$ at time t, W is a random variable (independent of the cell genome's IS count), and $\mathbf{v} = (v_1, \ldots, v_l)$ is the scaled left eigenvector to the eigenvalue λ_0 of the infinitesimal generator A [Athreya and Ney, 1972, p. 206]. Therefore, if $\mathbf{v}$ is rescaled so that $\sum_{k=1}^{l} v_k = 1$, its components $v_1, \ldots, v_l$ denote the limit distribution of IS counts in infected cells.

Figure 8 shows the computed limit distributions of IS counts per genome as a function of the HGT rate, for different parameter combinations. These limit distributions are approached asymptotically after the first IS infection occurred.

[insert figure 8 here]

Figure 8 shows that for the broad parameter range used in our model, most infected cells contain only one IS. The decrease in the fraction of cells with two, three, or more ISs per genome gets even steeper for higher HGT rates. This result can be understood by noting that the IS count distribution in our multi-type model is determined by the replicative transposition rate u opposing the fitness cost s per IS and the HGT rate h (the excision rate e is too small to be of any importance). As $h > s$ is necessary for a persisting infection (see subsection 3.1), we can distinguish between three scenarios: $u > h > s$, $h > u > s$, and $h > s > u$. In the first scenario $u > h > s$, replicative transposition increases the IS count of cells faster than new cells can be infected with one IS. Therefore, the IS count distribution gets shifted towards higher values until an equilibrium is reached with the increasing total fitness cost per cell. In the second and third scenarios $h > u > s$ or $h > s > u$, HGT infects new cells faster with one IS than the IS count of already infected cells can increase. Therefore, the IS count distribution is strongly L-shaped. Considering our model parameter range, the latter two scenarios are more probable, and therefore, the IS count distribution in our model is generally L-shaped. Because $h > s$ is a necessary condition for IS infection persistence, no IS count distribution can be shown in figure 8 for $h < s$. In fact, an infection can become extinct with certainty even for h slightly larger than s if the IS count distribution is no longer strictly dominated by cells with one IS (see the lower right graph in figure 8).

4 Discussion

An IS that provides a sufficiently large benefit to its host can rapidly rise to fixation through natural selection [Hall, 1999, Schneider and Lenski, 2004]. We are interested in the more challenging scenario, where an IS is slightly detrimental. When newly introduced into an uninfected host cell population, such an IS faces a situation analogous to that of a slightly detrimental mutant allele that has newly emerged in a population. Its frequency in the population is subject to random drift, and it is easily driven to extinction [Moran, 1962, Ohta, 1974, Kimura, 1983]. However, this analogy with population genetics is limited: most population-genetic models are

neither concerned with HGT, which can increase the number of cells carrying an IS for reasons different from selection and genetic drift, nor do they take into account the possibility of a genetic element increasing its number (and therefore its fitness cost) in a genome. Here, we focus on the interplay between HGT and other factors influencing the persistence of an infection with mobile genetic elements that can autonomously reproduce and increase their own number in an infected genome.

4.1 Survival probability

The linear dependency of p_{surv} on $h - s$ for low replicative transposition rate means that HGT stands in direct opposition to the selection against ISs. Specifically, an IS infection will only survive if the HGT rate h is higher than the fitness cost s of an IS. However, even if ISs have no fitness cost, the survival probability of an IS infection starting with one infected cell is small, because HGT rates are generally small. In a bulk environment (e.g. seawater), HGT rates are probably at most 10^{-5} to 10^{-4} events per infected cell and generation [Dahlberg et al., 1998]. This range of the HGT rate provides an upper bound for the difference $h - s$. Even for neutral ISs, the survival probability of an IS infection starting with one infected cell would therefore be 10^{-4} at most.

4.2 Time to extinction

The median time to extinction $T_{0,\text{med}} \approx 1 - (3h - s)/2$ is dominated by the comparatively large constant 1. This means that half of the IS infections that die out do so in merely one generation. However, the distribution of the time to extinction is highly right-skewed. Some infections can therefore survive for a much longer time before they eventually die out.

The relationship $T_{0,\text{med}} \approx 1 - (3h - s)/2$ seems paradoxical at first, as the median time to extinction decreases with *increasing* HGT rate and/or *decreasing* fitness cost. However, this is due to the following bias: we are examining only infections that become extinct, and with increasing HGT rate and/or decreasing fitness cost, populations of infected cells tend to spend less time lingering at low population sizes before they either die out or begin to grow. In other words, an infection's fate is determined more quickly for increased HGT rate and/or decreased fitness cost, thereby reducing the median time to extinction.

4.3 Time to threshold

The time to threshold can be very long, especially if the HGT rate is only slightly higher than the fitness cost and therefore their difference almost vanishes. For the upper bound $h \in [10^{-5}, 10^{-4}]$ gen.$^{-1}$ of $h - s$ we used before, the median time to reach a population size threshold of 10^8 infected cells is between 10^5 and $10^{5.8} = 6.3 \cdot 10^5$ generations (see figure 5). Generation times of bacteria living in the wild vary broadly, but with an assumed generation time of one day for *E. coli* [Gibbons and Kapsimalis, 1967, Savageau, 1983], the median time to threshold for these large HGT rates is between 300 and 1700 years. As the time to threshold is right-skewed, it can sometimes be much longer. Because our information about IS infections stems from limited samples, such long times to threshold would in practice make it difficult to detect many IS infections, even if they were successful in the end.

4.4 IS count distribution

Within broad parameter ranges, our model predicts that a large majority of infected cells harbor only one IS per genome, and the fraction of cells with more than one IS drops quickly

with increasing IS count. This holds even more for high HGT rates. The predicted distribution, biased towards very low IS counts, is corroborated by empirical data from more than 700 genomes, and it has also been observed in previous work based on a smaller number of genomes [Sawyer et al., 1987, Touchon and Rocha, 2007].

If the fitness cost is larger than the replicative transposition rate, the IS count distribution is highly skewed over the whole range of used HGT rates, with most cells harboring only one IS (see figure 8). To get an IS count distribution similar to the empirical distribution shown in figure 7, the fitness cost probably has to be somewhat smaller than the replicative transposition rate. The replicative transposition rate, in turn, is very low. We assume it to be in the interval $u \in [10^{-9}, 10^{-6}]$ gen.$^{-1}$. Our models therefore suggest that ISs might be effectively neutral in their effects on the host cell.

4.5 Effects of nonconstant HGT and transposition rates

In our model, we assume the replicative transposition rate and the HGT rate to be independent of the cell's IS count. We now discuss an alternative scenario, where the replicative transposition rate and/or the HGT rate linearly increase with the cell's IS count. Specifically, we discuss the effects of these scenarios on the IS count distribution, the survival probability of an IS infection, and the time to threshold. We do not discuss the effects on the extinction probability of an IS infection, because extinction happens fast and does not leave much time for transposition and HGT, and because our birth-and-death process model does not include transposition.

If the replicative transposition rate linearly increases with the IS count, the balance of forces determining the IS count distribution shifts: replicative transposition is strengthened in its opposition against fitness cost and HGT. Infected cells reach higher IS counts than if the replicative transposition rate is constant; although in most cases, the IS count distribution is still dominated by cells with one or a few ISs. Only if the replicative transposition rate is larger than the HGT rate (and therefore larger than the fitness cost), then the IS count distribution is dominated by cells with the highest IS count allowed in the model. This is an unrealistic scenario and not consistent with the observed IS count distribution. A shift towards higher IS counts increases the fitness cost and therefore reduces the survival probability; although only slightly so, as long as the IS count distribution is still dominated by cells with one or only a few ISs. For the same reason, the time to threshold does not noticeably change (but remember that for the time to threshold, we have to restrict our model to a maximum of $l = 5$ ISs per cell).

If the HGT rate linearly increases with the IS count, the IS count distribution shifts towards lower values, as more cells get infected with one IS. Together, the higher infection rate and the lower fitness cost induced by only one IS increase the survival probability of an infection, especially for HGT rates only slightly larger than the fitness cost of an IS. A higher infection rate and a lower fitness cost also slightly decrease the time to threshold.

If both the replicative transposition rate and the HGT rate increase linearly with the IS count, two opposing forces in shaping the IS count distribution are strengthened: infected cells will reach higher IS counts, and at the same time, cells with higher IS counts will infect more cells with only one IS. The IS count distribution then shifts towards higher IS counts, but less so than when only the replicative transposition rate linearly increases. The survival probability, on the other hand, is similar to the one observed when only the HGT rate linearly increases: although cells with higher IS counts bear a higher fitness cost, they also infect more cells with an IS and keep the IS infection spreading. For this reason, the time to threshold is also slightly lower than with constant replicative transposition and HGT rate, albeit not as low as when only the HGT rate linearly increases with the IS count.

4.6 Caveats

We here discuss the limitations of our analysis, some of which are caused by our model assumptions, whereas others are caused by limited data.

First, in our branching process models, we assume a well-mixed environment, where infected cells are surrounded by uninfected cells and where they are not clustered. The models are therefore not valid for bacteria living in a spatially structured environment, e.g. in a biofilm. Second, we assume that an infection starts with one cell that is infected with one IS. We note that in naturally occurring bacterial populations, the prevalence of infected cells is low (see [Wagner, 2006, Touchon and Rocha, 2007] and figure 7). Therefore, even if many new bacterial cells are introduced into an uninfected host cell population, probably only a few of these new cells are infected. This justifies our assumption. Third, we restrict HGT to transferring an IS copy only into uninfected cells. Again, this is no serious restriction: first, we only consider the early phase of an IS infection, with a low number of infected cells, and second, we assume infected cells to be well-mixed with and surrounded by uninfected cells, so that HGT into already infected cells can be neglected.

5 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.

MB thanks Dominik Heinzmann for many fruitful discussions about mathematical models, Nicole de la Chaux for her help with programming, and Corina Bichsel for her editorial help. He also thanks the reviewers for their helpful comments and suggestions.

6 Figures

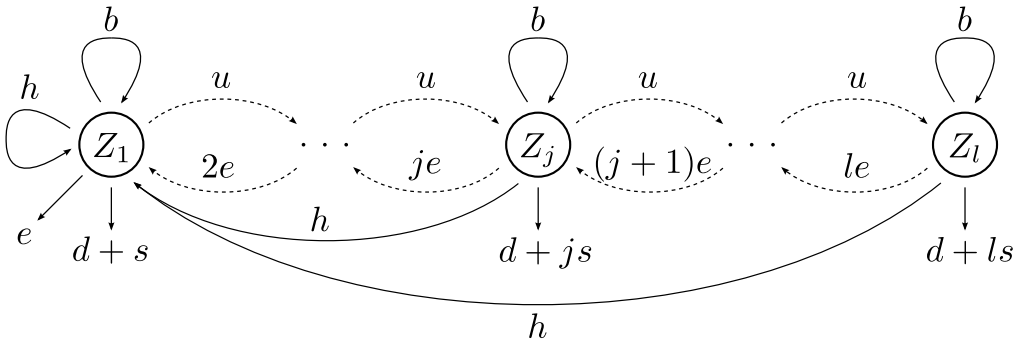

Figure 1: Multi-type model design. Z_k = number of cells with k ISs ($k \in \{1, \ldots, l\}$), b = birth rate per cell, d = base death rate per cell, u = replicative IS transposition rate per cell, e = IS excision rate per IS, h = HGT rate per cell, s = fitness cost per IS, and l = maximal IS count per genome (all rates are per host cell generation). Solid arrows indicate a change of total IS count and total infected cell count. Dashed arrows indicate a change of total IS count only.

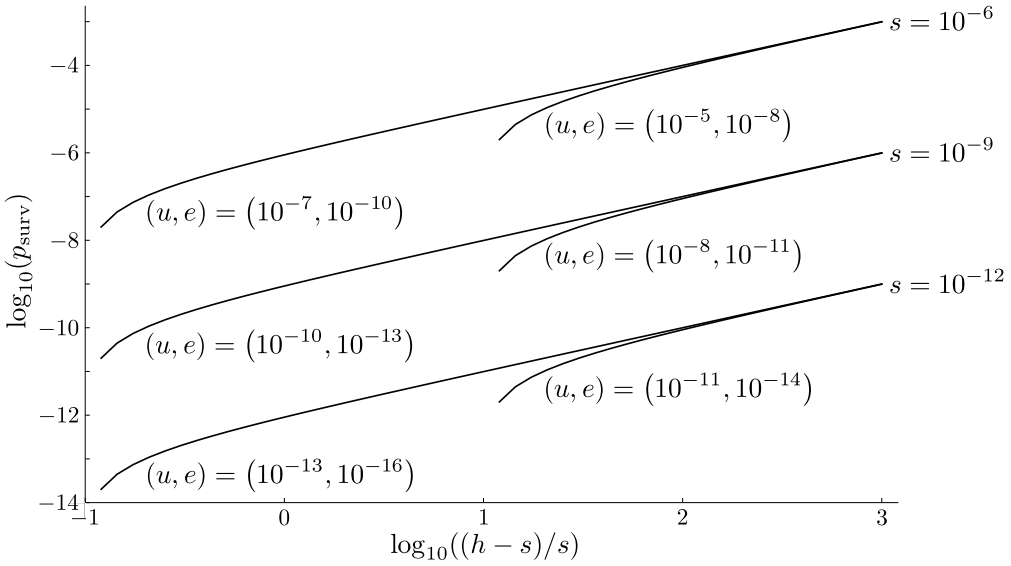

Figure 2: Computed survival probability p_{surv} of a population of infected cells, starting with one cell infected with a single IS, as a function of the relative difference $(h - s)/s$ between the HGT rate h and the fitness cost s, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s, u, e) \in \{(10^{-12}, 10^{-13}, 10^{-16}), (10^{-12}, 10^{-11}, 10^{-14}), (10^{-9}, 10^{-10}, 10^{-13}), (10^{-9}, 10^{-8}, 10^{-11}), (10^{-6}, 10^{-7}, 10^{-10}), (10^{-6}, 10^{-5}, 10^{-8})\}$ gen.$^{-1}$, $l = 50$.

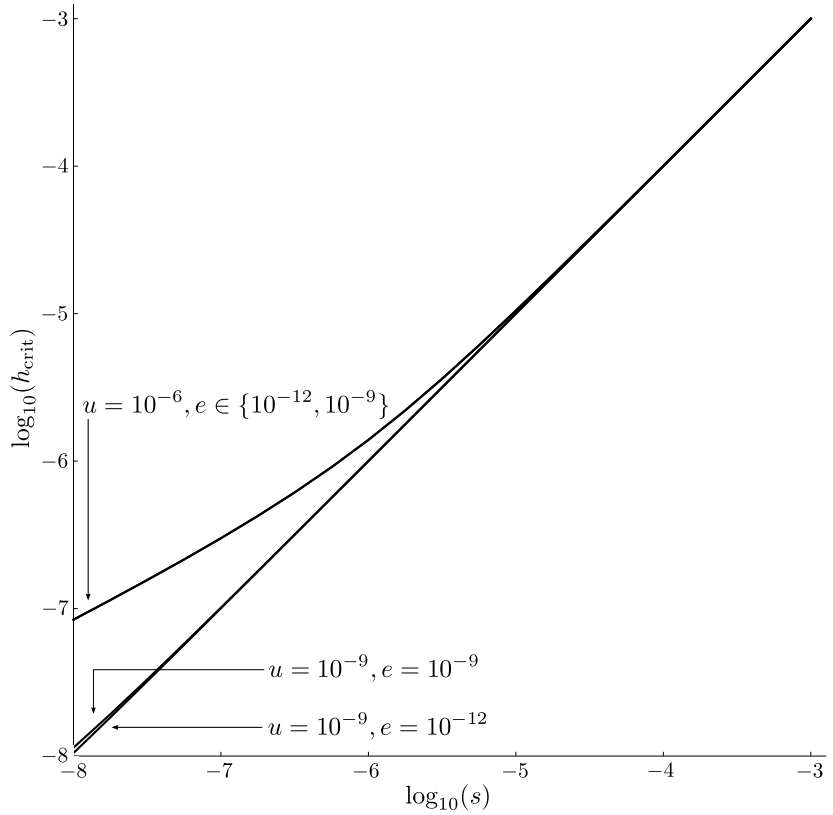

Figure 3: Computed critical HGT rate h_{crit} as a function of the fitness cost s, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(u, e) \in \left\{\left(10^{-9}, 10^{-12}\right), \left(10^{-9}, 10^{-9}\right), \left(10^{-6}, 10^{-12}\right), \left(10^{-6}, 10^{-9}\right)\right\}$ gen.$^{-1}$, $l = 50$.

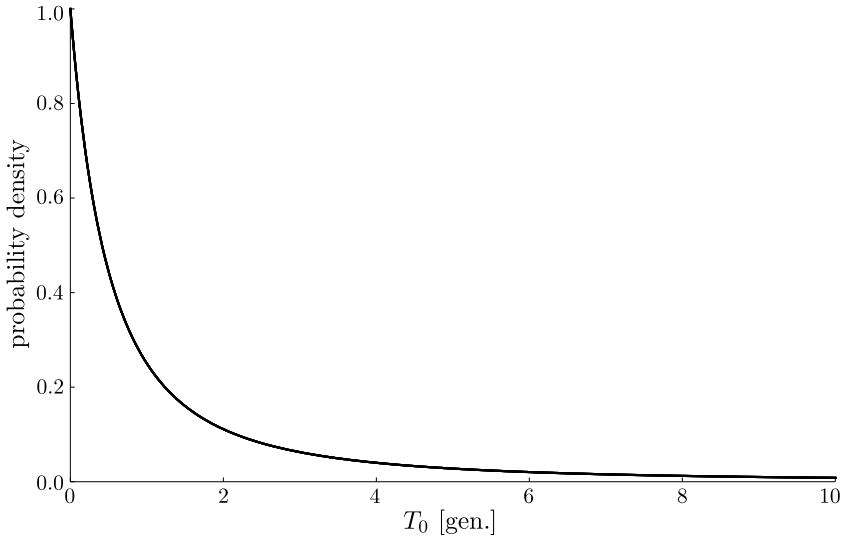

Figure 4: Probability density function of the time to extinction T_0, for different parameter combinations. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s, h) \in \left\{\left(10^{-12}, 10^{-7}\right), \left(10^{-12}, 10^{-4}\right), \left(10^{-6}, 10^{-4}\right)\right\}$ gen.$^{-1}$. The single line is an overlay of the graphs obtained when using the three parameter value combinations of s and h indicated above.

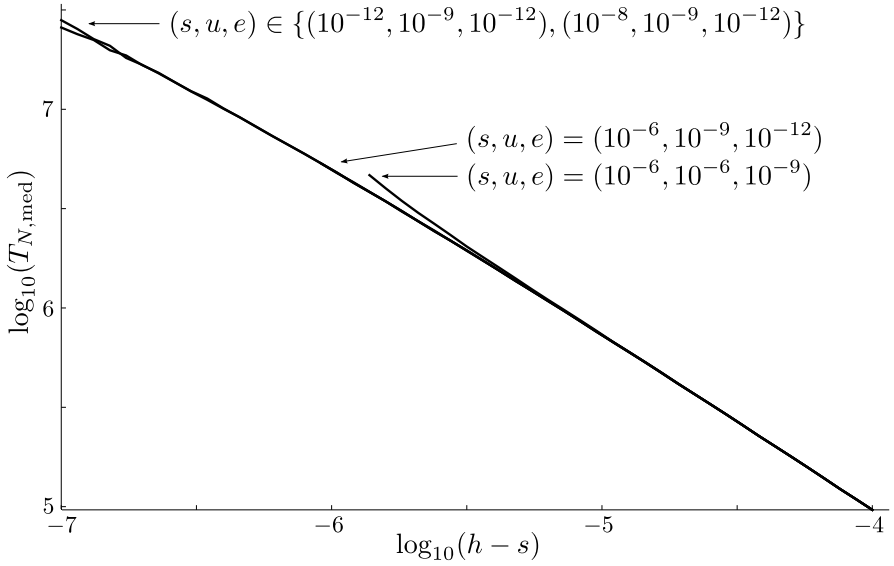

Figure 5: Computed median time $T_{N,\mathrm{med}}$ to a threshold of $N = 10^8$ infected cells as a function of the difference $h - s$ between the HGT rate and the fitness cost, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s,u,e) \in \left\{\left(10^{-12},10^{-9},10^{-12}\right),\ \left(10^{-8},10^{-9},10^{-12}\right),\left(10^{-6},10^{-9},10^{-12}\right),\left(10^{-6},10^{-6},10^{-9}\right)\right\}$ gen.$^{-1}$, $l = 5$. Because computing the characteristic function is feasible only for moderate fitness costs, not all graphs extend to the full range of the difference between the HGT rate and the fitness cost. The arrows mark the beginnings of the curves with the corresponding parameter sets.

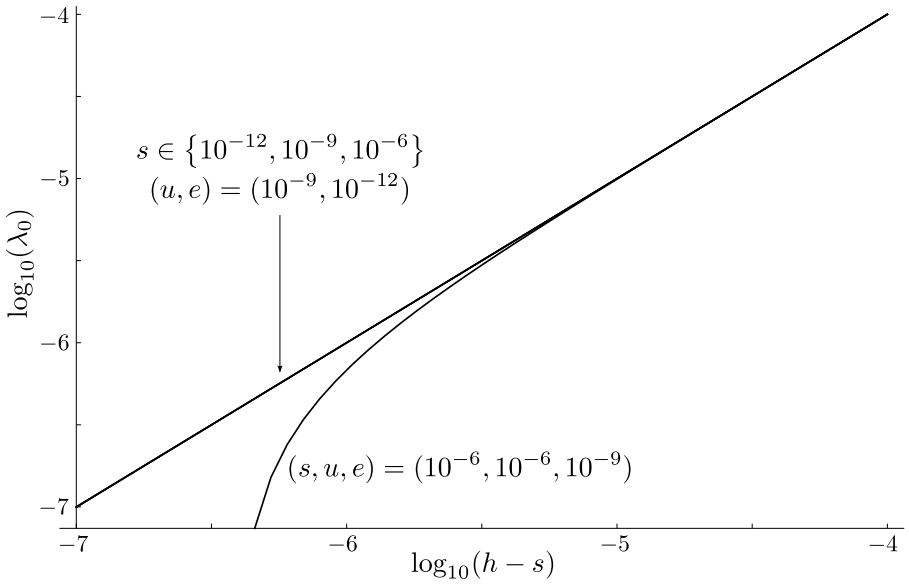

Figure 6: Growth rate λ_0 as a function of the difference $h - s$ between the HGT rate and the fitness cost, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s,u,e) \in \left\{\left(10^{-12},10^{-9},10^{-12}\right),\left(10^{-8},10^{-9},10^{-12}\right),\left(10^{-6},10^{-9},10^{-12}\right),\left(10^{-6},10^{-6},10^{-9}\right)\right\}$ gen.$^{-1}$, $l = 50$.

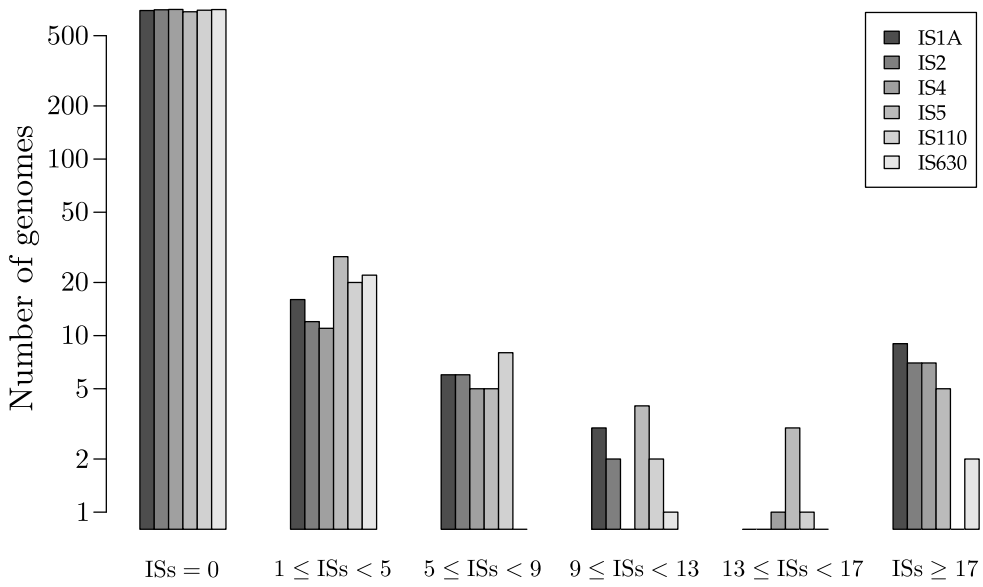

Figure 7: IS count distribution of the six most abundant ISs in 728 fully sequenced bacterial genomes (June 2009). "ISs" means "IS count per genome". Note the logarithmic vertical axis.

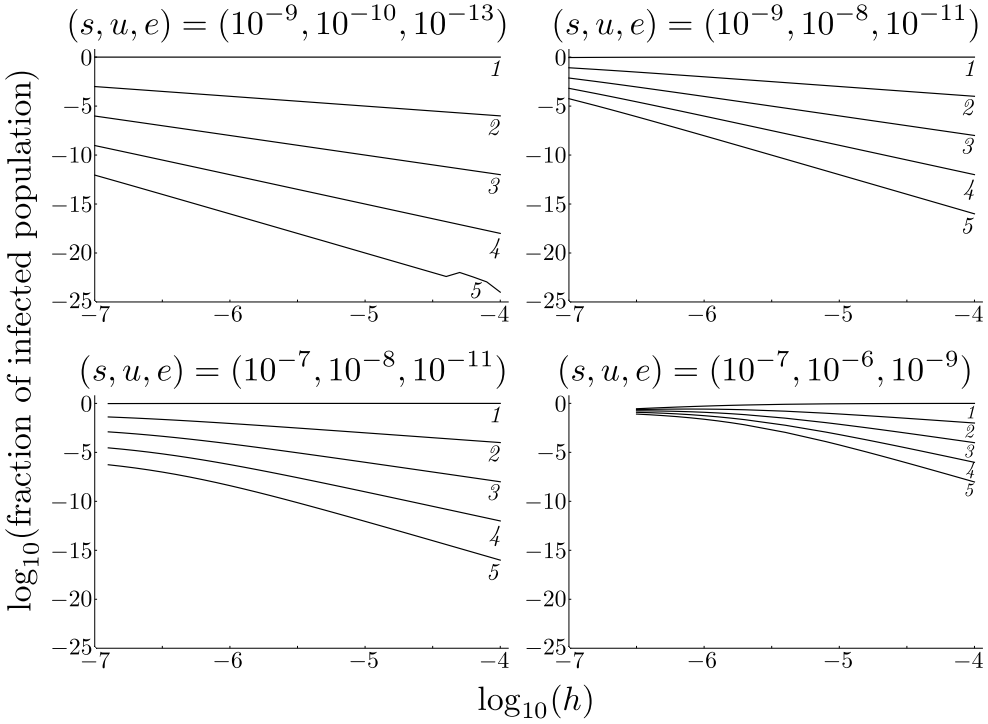

Figure 8: Computed IS count distribution as a function of the HGT rate h, for different parameter combinations. Note the logarithmic scales. Parameter values: $b = d = 1$ gen.$^{-1}$, $(s,u,e) \in \left\{\left(10^{-9},10^{-10},10^{-13}\right), \left(10^{-9},10^{-8},10^{-11}\right), \left(10^{-7},10^{-8},10^{-11}\right), \left(10^{-7},10^{-6},10^{-9}\right)\right\}$ gen.$^{-1}$, $l = 50$ (but at most 5 ISs per infected cell are shown). The numbers in italics indicate the IS count per genome.

7 Tables

Table 1 - Rates

Event	**Reported rates**		**Model rates**	
Transposition	Conservative	10^{-7} – 10^{-4}	Replicative	10^{-9} – 10^{-6}
Excision		10^{-10}		10^{-12} – 10^{-9}
HGT	Transduction	10^{-8}		
	Conjugation	10^{-6} – 10^{-5}	Total	10^{-7} – 10^{-4}
	Transformation	10^{-6} – 10^{-3}		
Fitness cost		–		10^{-12} – 10^{-6}

Event rates reported by different authors (rates are converted into events per cell or IS and day), and corresponding parameter ranges used in our models. Model rates are per cell and cell generation or, in the case of the fitness cost, per IS and cell generation. Origin of reported rates: conservative transposition [Kleckner, 1989, Chandler and Mahillon, 2002], excision [Kleckner, 1989], transduction [Jiang and Paul, 1998], conjugation [Dahlberg et al., 1998], transformation [Williams et al., 1996].

A Models: Multi-type model

The probability generating function of a multi-type branching process is defined as

$$\begin{aligned}\mathbf{g}(\mathbf{z}) &= \sum_{\mathbf{j}} \mathbf{p}(\mathbf{j})\mathbf{z}^{\mathbf{j}} \\ &= \left(\sum_{(j_{11},\ldots,j_{1l})} p_1(j_{11},\ldots,j_{1l}) z_1^{j_{11}} \cdot \ldots \cdot z_l^{j_{1l}}, \ldots, \sum_{(j_{l1},\ldots,j_{ll})} p_l(j_{l1},\ldots,j_{ll}) z_1^{j_{l1}} \cdot \ldots \cdot z_l^{j_{ll}} \right),\end{aligned}$$

where $p_k(j_{k1},\ldots,j_{kl})$ is the probability of a particle of type k (here: a cell with k ISs) to produce $j_{k1},\ldots,j_{kl}$ particles of type $1,\ldots,l$. In our case, we get the following probability generating function:

$$\begin{aligned}g_1(\mathbf{z}) &= \frac{b+h}{a_1} z_1^2 + \frac{d+s+e}{a_1} + \frac{u}{a_1} z_2 \\ g_j(\mathbf{z}) &= \frac{b}{a_j} z_j^2 + \frac{d+js}{a_j} + \frac{u}{a_j} z_{j+1} + \frac{je}{a_j} z_{j-1} + \frac{h}{a_j} z_1 z_j \quad (1 < j < l) \\ g_l(\mathbf{z}) &= \frac{b}{a_l} z_l^2 + \frac{d+ls}{a_l} + \frac{le}{a_l} z_{l-1} + \frac{h}{a_l} z_1 z_l,\end{aligned}$$

where $a_k = b + d + ks + u + ke + h$ is the event rate of a cell with k ISs (see subsection 3.1).

From the probability generating function, we derive the infinitesimal generating function $\tilde{g}_k(\mathbf{z}) = a_k(g_k(\mathbf{z}) - z_k)$:

$$\begin{aligned}\tilde{g}_1(\mathbf{z}) &= (b+h)z_1^2 - (b+h+d+s+u+e)z_1 + uz_2 + d + s + e \\ \tilde{g}_j(\mathbf{z}) &= bz_j^2 - (b+h+d+js+u+je)z_j + uz_{j+1} + jez_{j-1} + hz_1z_j + d + js \\ \tilde{g}_l(\mathbf{z}) &= bz_l^2 - (b+h+d+ls+le)z_l + lez_{l-1} + hz_1z_l + d + ls\end{aligned}$$

and the infinitesimal generator $A = (a_{ij}) = a_i b_{ij}$, where $b_{ij} = \left.\frac{\partial g_i(\mathbf{z})}{\partial z_j}\right|_{\mathbf{z}=\mathbf{1}} - \delta_{ij}$:

$$A = \begin{pmatrix} b+h-d-s-u-e & u & & & & & & \\ h+2e & b-d-2s-u-2e & u & & & & & \\ h & 3e & b-d-3s-u-3e & u & & & & \\ \vdots & & & \ddots & & & & \\ h & & & je & b-d-js-u-je & u & & \\ \vdots & & & & & \ddots & & \\ h & & & & & le & b-d-ls-le \end{pmatrix}$$

B Results: Time to threshold

To extend the ordinary differential equations given in subsection 3.3 to $x = 0$, we observe that

$$\begin{aligned}\frac{d\varphi^m(x)}{dx} &= \frac{d}{dx}\mathbb{E}\left(e^{iW^m x}\right) = \frac{d}{dx}\int_0^\infty e^{itx} f^m(t)dt \overset{\text{Leibniz}}{=} \int_0^\infty \frac{\partial}{\partial x}\left[e^{itx} f^m(t)\right] dt \\ &= \int_0^\infty ite^{itx} f^m(t)dt = \mathbb{E}\left(iW^m e^{iW^m x}\right),\end{aligned}$$

where $f^m(t)$ is the probability distribution of W^m, and so $\left.\frac{d\varphi^m(x)}{dx}\right|_{x=0} = i\mathbb{E}\left(W^m\right) = iu_m$, where u_m is the m-th component of the scaled right eigenvector $\mathbf{u}$ to the eigenvalue λ_0 of the infinitesimal generator A.

Therefore, the ordinary differential equation system for $\varphi^m(x)$, $m \in \{1, \ldots, l\}$, is

$$\begin{aligned}
\frac{d\varphi^1(x)}{dx} &= \frac{1}{\lambda_0 x}\,[\,(b+h)(\varphi^1(x))^2 - (b+h+d+s+u+e)\varphi^1(x) \\
&\qquad + u\varphi^2(x) + d + s + e\,] \\
\frac{d\varphi^j(x)}{dx} &= \frac{1}{\lambda_0 x}\,[\,h\varphi^1(x)\varphi^j(x) + b(\varphi^j(x))^2 - (b+h+d+js+u+je)\varphi^j(x) \\
&\qquad + u\varphi^{j+1}(x) + je\varphi^{j-1}(x) + d + js\,] \quad (1 < j < l) \\
\frac{d\varphi^l(x)}{dx} &= \frac{1}{\lambda_0 x}\,[\,h\varphi^1(x)\varphi^l(x) + b(\varphi^l(x))^2 - (b+h+d+ls+le)\varphi^l(x) \\
&\qquad + le\varphi^{l-1}(x) + d + ls\,]
\end{aligned}$$

if $x \neq 0$, and

$$\left.\frac{d\varphi^m(x)}{dx}\right|_{x=0} = iu_m \quad \text{for } m \in \{1, \ldots, l\}$$

if $x = 0$, with

$$\varphi^m(0) = 1 \text{ for } m \in \{1, \ldots, l\}.$$

References

[Athreya and Ney, 1972] Athreya, K. B. and Ney, P. E. (1972). *Branching Processes*. Springer Verlag, Berlin.

[Basten and Moody, 1991] Basten, C. J. and Moody, M. E. (1991). A branching-process model for the evolution of transposable elements incorporating selection. *Journal of Mathematical Biology*, 29:743–761.

[Berg, 1977] Berg, D. E. (1977). Insertion and excision of the transposable kanamycin resistance determinant Tn5. In Bukhari, A. I., Shapiro, J. A., and Adhya, S. L., editors, *DNA insertion elements, plasmids, and episomes*, pages 205–212. Cold Spring Harbor Laboratory.

[Berg, 1989] Berg, D. E. (1989). Transposon Tn5. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 185–210. American Society for Microbiology, Washington, D.C.

[Bergthorsson and Ochman, 1998] Bergthorsson, U. and Ochman, H. (1998). Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Molecular Biology and Evolution*, 15(1):6–16.

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

[Chandler and Mahillon, 2002] Chandler, M. and Mahillon, J. (2002). Insertion sequences revisited. In Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M., editors, *Mobile DNA II*, pages 305–366. American Society for Microbiology, Washington, D.C.

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

[Dahlberg et al., 1998] Dahlberg, C., Bergström, M., and Hermansson, M. (1998). In situ detection of high levels of horizontal plasmid transfer in marine bacterial communities. *Applied and Environmental Microbiology*, 64(7):2670–2675.

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

[Doolittle and Sapienza, 1980] Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284:601–603.

[Dröge et al., 1999] Dröge, M., Pühler, A., and Selbitschka, W. (1999). Horizontal gene transfer among bacteria in terrestrial and aquatic habitats as assessed by microcosm and field studies. *Biology and Fertility of Soils*, 29(3):221–245.

[Egner and Berg, 1981] Egner, C. and Berg, D. E. (1981). Excision of transposon Tn5 is dependent on the inverted repeats but not on the transposase function of Tn5. *Proceedings of the National Academy of Sciences of the United States of America*, 78(1):459–463.

[Feller, 1939] Feller, W. (1939). Die Grundlagen der Volterraschen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheoretica*, 5(1):11–40.

[Fisher, 1922] Fisher, R. A. (1922). On the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42:321–341.

[Galas and Chandler, 1989] Galas, D. J. and Chandler, M. (1989). Bacterial insertion sequences. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 109–162. American Society for Microbiology, Washington, D.C.

[Gibbons and Kapsimalis, 1967] Gibbons, R. J. and Kapsimalis, B. (1967). Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *Journal of Bacteriology*, 93(1):510–512.

[Gogarten and Townsend, 2005] Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687.

[Haccou et al., 2005] Haccou, P., Jagers, P., and Vatutin, V. A. (2005). *Branching Processes: Variation, Growth, and Extinction of Populations*. Cambridge University Press, New York.

[Haldane, 1927] Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection, part V: selection and mutation. *Proceedings of the Cambridge Philosophical Society*, 23:838–844.

[Hall, 1999] Hall, B. G. (1999). Transposable elements as activators of cryptic genes in E. coli. *Genetica*, 107:181–187.

[Harris, 1951] Harris, T. E. (1951). Some mathematical models for branching processes. In Neyman, J., editor, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 305–328. University of California Press, Berkeley, California.

[Jagers, 1975] Jagers, P. (1975). *Branching Processes with Biological Applications*. Wiley, London.

[Jiang and Paul, 1998] Jiang, S. C. and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology*, 64(8):2780–2787.

[Johnson et al., 1995] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2. John Wiley and Sons, Inc., New York, 2nd edition.

[Kendall, 1948] Kendall, D. G. (1948). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, 19(1):1–15.

[Kimmel and Axelrod, 2002] Kimmel, M. and Axelrod, D. E. (2002). *Branching Processes in Biology*. Springer-Verlag New York, Inc.

[Kimura, 1983] Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

[Kleckner, 1989] Kleckner, N. (1989). Transposon Tn10. In Berg, D. E. and Howe, M. M., editors, *Mobile DNA*, pages 227–268. American Society for Microbiology, Washington, D.C.

[Mahillon et al., 2009] Mahillon, J., Siguier, P., and Chandler, M. (2009). IS Finder. `http://www-is.biotoul.fr`.

[Mayaux et al., 1984] Mayaux, J.-F., Springer, M., Graffe, M., Fromant, M., and Fayat, G. (1984). Is*4* transposition in the attenuator region of the *Escherichia coli pheS,T* operon. *Gene*, 30:137–146.

[McClintock, 1950] McClintock, B. (1950). The origin and behaviour of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–355.

[Moody, 1988] Moody, M. E. (1988). A branching-process model for the evolution of transposable elements. *Journal of Mathematical Biology*, 26:347–357.

[Moran, 1962] Moran, P. A. P. (1962). *The statistical processes of evolutionary theory*. Oxford University Press.

[NCBI, 2009] NCBI (2009). National Center for Biotechnology Information. `http://www.ncbi.nlm.nih.gov`.

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

[Ohta, 1974] Ohta, T. (1974). Mutational pressure as main cause of molecular evolution and polymorphism. *Nature*, 252:351–354.

[Orgel and Crick, 1980] Orgel, L. E. and Crick, F. H. C. (1980). Selfish DNA: the ultimate parasite. *Nature*, 284:604–607.

[Powell, 1955] Powell, E. O. (1955). Some features of the generation times of individual bacteria. *Biometrika*, 42(1/2):16–44.

[R Development Core Team, 2008] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Savageau, 1983] Savageau, M. A. (1983). *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *The American Naturalist*, 122(6):732–744.

[Sawyer et al., 1987] Sawyer, S. A., Dykhuizen, D. E., DuBose, R. F., Green, L., Mutangadura-Mhlanga, T., Wolczyk, D. F., and Hartl, D. L. (1987). Distribution and Abundance of Insertion Sequences Among Natural Isolates of Escherichia coli. *Genetics*, 115:51–63.

[Schneider and Lenski, 2004] Schneider, D. and Lenski, R. E. (2004). Dynamics of insertion sequence elements during experimental evolution of bacteria. *Research in Microbiology*, 155:319–327.

[Sewastjanow, 1975] Sewastjanow, B. A. (1975). *Verzweigungsprozesse*. Akademie Verlag, Berlin.

[Shapiro, 1999] Shapiro, J. A. (1999). Transposable elements as the key to a 21st century view of evolution. *Genetica*, 107:171–179.

[So and McCarthy, 1980] So, M. and McCarthy, B. J. (1980). Nucleotide sequence of the bacterial transposon TN1681 encoding a heat-stable (ST) toxin and its identification in enterotoxigenic Escherichia coli strains. *Proceedings of the National Academy of Sciences of the United States of America*, 77(7):4011–4015.

[Sørensen et al., 2005] Sørensen, S. J., Bailey, M., Hansen, L. H., Kroer, N., and Wuertz, S. (2005). Studying Plasmid Horizontal Transfer *in situ*: a Critical Review. *Nature Reviews Microbiology*, 3(9):700–710.

[Tavakoli and Derbyshire, 2001] Tavakoli, N. P. and Derbyshire, K. M. (2001). Tipping the balance between replicative and simple transposition. *The EMBO Journal*, 20(11):2923–2930.

[Thompson et al., 2004] Thompson, J. R., Randa, M. A., Marcelino, L. A., Tomita-Michell, A., Lim, E., and Polz, M. F. (2004). Diversity and Dynamics of a North Atlantic Coastal *Vibrio* Community. *Applied and Environmental Microbiology*, 70(7):4103–4110.

[Top and Springael, 2003] Top, E. M. and Springael, D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Current Opinion in Biotechnology*, 14:262–269.

[Touchon and Rocha, 2007] Touchon, M. and Rocha, E. P. C. (2007). Causes of insertion sequences abundance in prokaryotic genomes. *Molecular Biology and Evolution*, 24(4):969–981.

[Wagner, 2006] Wagner, A. (2006). Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes. *Molecular Biology and Evolution*, 23(4):723–733.

[Wagner et al., 2007] Wagner, A., Lewis, C., and Bichsel, M. (2007). A survey of bacterial insertion sequences using IScan. *Nucleic Acids Research*, 35(16):5284–5293.

[Williams et al., 1996] Williams, H. G., Day, M. J., Fry, J. C., and Stewart, G. J. (1996). Natural transformation in river epilithon. *Applied and Environmental Microbiology*, 62(8):2994–2998.

[Zerbib et al., 1985] Zerbib, D., Gamas, P., Chandler, M., Prentki, P., Bass, S., and Galas, D. (1985). Specificity of insertion of IS*1*. *Journal of Molecular Biology*, 185:517–524.