Supplementary information

This supplementary information contains the following material:

1) A detailed description of the models and methods used (p.1)
2) An additional table (Table S1) and figure (Figure S1) illustrating the model (p.6)
3) Three figures (S2 to S4) showing additional results of the Basic Model (p.7)
4) Two figures (S5 and S6) showing additional results of Model Extension 1 (p.9)
5) Two figures (S7 and S8) showing additional results of Model Extension 2 (p.10)

Methods

General population setup

We consider a population of two competing types of bacteria: those carrying a functional integrase gene and those with a non-functional integrase gene. The number of bacteria of each type will be denoted by *X* and *Y*, respectively, and *N* is the total population size. Bacteria are further characterized by their integron cassette genes. We assume that there *n* different such genes, so when allowing for duplications of cassettes within the integron and considering different integron lengths up to a maximum number of *k* cassettes, there are a total of $1 + n + n^2 + \cdots + n^k = (n^{k+1} - 1)/(n - 1)$ different integron genotypes. (The number 1 in this formula represents an "empty" integron without any gene cassettes.) The total set of possible integron genotypes is denoted by *G*.

The bacteria exhibit logistic growth with a maximum growth rate standardized to a value of 1 and a carrying capacity *K*. The bacteria also die at a baseline death rate η_0. This death rate may be further increased by an amount η_I in individuals carrying a functional integrase (cost of integrase) and by an additional, stress-induced death rate that depends on both genotype and environmental conditions (see below). At a rate μ, mutations in the integrase gene lead to a non-functional version of that gene. For simplicity, we assume that there is no back mutation from a non-functional to a functional integrase. This assumption seems justified because the rate of back mutation is expected to be much smaller than μ, as disrupting gene function is much more likely to occur than reversal of that disruption. In test simulations where back mutations did occur, this seemed to have little effect on the outcome (compare Figures 3 and S3).

Integrase action

Bacteria with a functional integrase undergo reshuffling of their gene cassettes at a rate ρ. At any such an event, a single, randomly chosen cassette is excised from the integron. With probability θ, the excised cassette is re-inserted in the first position of the integron, resulting in a new permutation of gene order. Conversely, with probability $1 - \theta$, the excised gene cassette is lost, resulting in a shorter integron (see Fig. 1B). Note that although insertion of cassettes in positions other than the first position within an integron have been reported, insertion in the first position is generally favoured (Collis et al., 1993, Collis et al.,

2001). To keep our model as simple as possible, we therefore assume that all cassette insertions occur in the first integron position.

Mathematically, integrons are described as vectors **g** of length *k* with entries 0, 1, 2, …, *n*, where 0 indicates the absence of cassettes at the end of the integron array. Integrase activity then converts one such vector into another, with probabilities given by two matrices $\mathbf{M}^{\text{exc}}$ (for excision of cassettes only) and $\mathbf{M}^{\text{int}}$ (for excision followed by re-integration). Table S1 gives these two matrices for the simple case of only two cassette genes; for the case $n = k = 3$ that will generally be assumed, these two matrices have dimension 40x40.

Stress-induced death and gene expression

We assume varying environmental conditions involving a number of stressors (e.g., presence of different antibiotics) that increase the death rate. At any given time point one or several of these stressors may be present in the population, as specified by a vector $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots S_n(t))$, where the *i*th component indicates the presence ($S_i = 1$) or absence ($S_i = 0$) of stressor *i*. Stressors appear and disappear randomly and independently from each other over time, following a continuous-time Markov process with the following transition rate matrix between the two states 0 and 1:

$$\frac{\sigma_{\text{vel}}}{2}\begin{pmatrix} -\frac{1}{1-\sigma_{\text{mean}}} & \frac{1}{1-\sigma_{\text{mean}}} \\ \frac{1}{\sigma_{\text{mean}}} & -\frac{1}{\sigma_{\text{mean}}} \end{pmatrix}. \quad (1)$$

This matrix is parameterized such that σ_{mean} gives the average fraction of time that any given stressor is present whilst σ_{vel} gives the velocity of switching between presence and absence of that stressor ($1/\sigma_{\text{vel}}$ is the average time between switching events).

Each of the cassettes provides resistance to one stressor (with corresponding index), but this resistance is a decreasing function of expression levels. We assume that there exists only a single promoter for all gene cassettes. The first cassette is fully expressed (normalized to expression level 1), but expression of the other gene cassettes is assumed to decline exponentially with increasing distance from the promoter, consistent with *attC* sites acting as translation control elements or as weak transcription terminators (Collis & Hall, 1995, Jacquier et al., 2009). Expression level of a cassette in the *i*th position in the integron is thus given by $E_i = e^{-\beta(i-1)}$, where β is a parameter determining how fast expression decreases as a function of position within the gene cassette (illustrated in Fig. S1A). Total gene expression of a specific cassette of type *j* in an integron genotype **g** (potentially containing duplicates of the cassette) is then obtained as the sum of all individual gene expressions, or $E_{\mathbf{g},j}^{\text{total}} = \sum_{i=1}^{k} \delta_{j,g_i} E_i$. Here, the Kronecker delta δ_{j,g_i} indicates whether the cassette in position *i* of the integron is of type *j* ($\delta_{j,g_i} = 1$) or not ($\delta_{j,g_i} = 0$).

Overall, the stress-induced death rate of integron genotype **g** at any given time is given by the function

$$h_{\mathbf{g}}(t) = \sum_{j=1}^{n} S_j(t)\, \eta_S \exp\left(-\gamma E_{\mathbf{g},j}^{\text{total}}\right). \quad (2)$$

The summation is over all stressors *j*, corresponding to all cassette alleles. This function assumes that each stressor causes an increase η_S in death rate, which can be reduced by expression of cassette genes. The parameter γ determines how strongly death rate declines with increasing total expression level of the corresponding cassette genes (see Fig. S1B). Figure 1C illustrates how the stress-induced death rate $h_{\mathbf{g}}$ depends on both the cassette array genotype and the presence of stressors in the environment.

Differential equations and simulation methods

Based on these above assumptions, we can write down a system of ordinary differential equations describing the evolutionary dynamics in the population:

$$\frac{\mathrm{d}X_{\mathbf{g}}}{\mathrm{d}t} = X_{\mathbf{g}}\left(1 - \frac{N}{K}\right) - (\eta_0 + \eta_I + h_{\mathbf{g}} + \rho + \mu)X_{\mathbf{g}} + \rho \sum_{j \in G} X_{\mathbf{j}}\left((1-\theta)M_{\mathbf{jg}}^{\text{exc}} + \theta M_{\mathbf{jg}}^{\text{int}}\right)$$

$$\frac{\mathrm{d}Y_{\mathbf{g}}}{\mathrm{d}t} = Y_{\mathbf{g}}\left(1 - \frac{N}{K}\right) - (\eta_0 + h_{\mathbf{g}})Y_{\mathbf{g}} + \mu X_{\mathbf{g}} \quad (3)$$

As described above, the stress-induced death rate $h_{\mathbf{g}}$ is not only genotype-dependent but also time-dependent in a stochastic manner; this means that mathematically speaking the model is a hybrid dynamical system with stochastic switching between different deterministic dynamics given by a system of ordinary differential equations. Note that for simplicity time dependency for $X_{\mathbf{g}}(t), Y_{\mathbf{g}}(t)$ and $h_{\mathbf{g}}(t)$ is not shown.

The model equations were implemented and numerically solved using the software package Mathematica version 10.0 (Wolfram Research, Inc.). Unless stated otherwise, all simulations were run for 10,000 time units and in 100 replicates; simulations were generally started from an initial population of 10^6 bacteria carrying a functional integrase and cassette genes 1, 2 and 3 in that order ($X_{\mathbf{123}}(0) = 10^6, X_{\mathbf{g}}(0) = 0$ for all other genotypes **g**, and $Y_{\mathbf{g}}(0) = 0$ for all genotypes **g**). All parameters of the model with their standard values are listed in Table 1. Parameter values chosen are based on experimental results where available (but see Discussion).

Model extension 1: stress-dependent integrase activity

To account for the finding that integrase expression can be induced by the SOS response pathway in some integrons (Guerin et al., 2009, Cambray et al., 2011), we extended our model so that the integrase is expressed only when the overall level of stress-induced death reaches a certain threshold value φ. Accordingly, reshuffling of the integron takes place and the integrase is costly only above certain levels of stress. The extended model can be written as

$$\begin{aligned}\frac{\mathrm{d}X_{\mathbf{g}}}{\mathrm{d}t} &= X_{\mathbf{g}}\left(1-\frac{N}{K}\right)-\left(\eta_0+\chi_\varphi(h_{\mathbf{g}})\eta_I+h_{\mathbf{g}}+\chi_\varphi(h_{\mathbf{g}})\rho+\mu\right)X_{\mathbf{g}}\\ &\quad +\rho\textstyle\sum_{\mathbf{j}\in G}\chi_\varphi(h_{\mathbf{j}})X_{\mathbf{j}}\left((1-\theta)M^{\mathsf{exc}}_{\mathbf{jg}}+\theta M^{\mathsf{int}}_{\mathbf{jg}}\right),\\ \frac{\mathrm{d}Y_{\mathbf{g}}}{\mathrm{d}t} &= Y_{\mathbf{g}}\left(1-\tfrac{N}{K}\right)-(\eta_0+h_{\mathbf{g}})Y_{\mathbf{g}}+\mu X_{\mathbf{g}},\end{aligned} \tag{4}$$

where χ_φ is a step-function defined as

$$\chi_\varphi(x)=\begin{cases}0 & \text{for } x<\varphi\\ 1 & \text{for } x\geq\varphi\end{cases}. \tag{5}$$

Model extension 2: horizontal gene transfer (HGT)

We also extended our model to incorporate HGT. Specifically, we assume that bacteria with a functional integrase pick up cassette genes from other bacteria (both with and without a functional integrase) and incorporate those cassettes in the first position of their integron. During this process, the recipient genotype is converted to a new genotype, but the donor remains unchanged. We assume a mass-action principle where the rate of HGT is proportional to the abundance of donor individuals. HGT takes place at a rate τ. The tensor **T** is used to describe genotypic conversions of a recipient strain following HGT; here, $T_{\mathbf{ijg}}$ is the fraction of bacteria of genotype **g** that arise from a donor with genotype **i** and a recipient with genotype **j**. For example, with a donor genotype **i**=12 and a recipient genotype **j**=10 (using the notation in Table S1), $T_{\mathbf{ijg}}$=1/2 for new genotypes **g**=11 and **g**=21, but $T_{\mathbf{ijg}}$=0 for all other genotypes **g**. When the donor does not carry any cassettes (**i**=00), the recipient is not altered (i.e., $T_{\mathbf{ijg}}$=1 for **j**=**g** and $T_{\mathbf{ijg}}$=0 otherwise).

The system of differential equations for this model is given by

$$\begin{aligned}\frac{\mathrm{d}X_{\mathbf{g}}}{\mathrm{d}t} &= X_{\mathbf{g}}\left(1-\frac{N}{K}\right)-\left(\eta_0+\eta_I+h_{\mathbf{g}}+\rho+\mu+\tau N\right)X_{\mathbf{g}}\\ &\quad +\rho\sum_{\mathbf{j}\in G}X_{\mathbf{j}}\left((1-\theta)M^{\mathsf{exc}}_{\mathbf{jg}}+\theta M^{\mathsf{int}}_{\mathbf{jg}}\right)+\tau\sum_{\mathbf{i},\mathbf{j}\in G}(X_{\mathbf{i}}+Y_{\mathbf{i}})X_{\mathbf{j}}T_{\mathbf{ijg}}\\ \frac{\mathrm{d}Y_{\mathbf{g}}}{\mathrm{d}t} &= Y_{\mathbf{g}}\left(1-\tfrac{N}{K}\right)-(\eta_0+h_{\mathbf{g}})Y_{\mathbf{g}}+\mu X_{\mathbf{g}}\end{aligned} \tag{6}$$

Note that in this model, the integrase still affects the gene order within the cassette array, but unlike in Model extension 1, is not stress-induced.

References:

Cambray, G., Sanchez-Alberola, N., Campoy, S., *et al.* (2011). Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. Mobile DNA 2: 6.

Collis, C. M., Grammaticopoulos, G., Briton, J., Stokes, H. W. & Hall, R. M. (1993). Site-specific insertion of gene cassettes into integrons. Mol Microbiol 9: 41-52.

Collis, C. M. & Hall, R. M. (1995). Expression of Antibiotic-Resistance Genes in the Integrated Cassettes of Integrons. Antimicrob. Agents Chemother 39: 155-162.

Collis, C. M., Recchia, G. D., Kim, M. J., Stokes, H. W. & Hall, R. M. (2001). Efficiency of recombination reactions catalyzed by class 1 integron integrase IntI1. J Bacteriol 183: 2535-2542.

Guerin, E., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S.*, et al.* (2009). The SOS Response Controls Integron Recombination. Science 324: 1034-1034.

Jacquier, H., Zaoui, C., Sanson-Le Pors, M. J., Mazel, D. & Bercot, B. (2009). Translation regulation of integrons gene cassette expression by the attC sites. Mol Microbiol 72: 1475-86.

Additional table and figures

Table S1. Example transition probabilities between integron genotypes.

$\mathbf{M}^{exc}$	00	10	20	11	12	21	22
00	1	0	0	0	0	0	0
10	1	0	0	0	0	0	0
20	1	0	0	0	0	0	0
11	0	1	0	0	0	0	0
12	0	1/2	1/2	0	0	0	0
21	0	1/2	1/2	0	0	0	0
22	0	0	1	0	0	0	0

$\mathbf{M}^{int}$	00	10	20	11	12	21	22
00	1	0	0	0	0	0	0
10	0	1	0	0	0	0	0
20	0	0	1	0	0	0	0
11	0	0	0	1	0	0	0
12	0	0	0	0	1/2	1/2	0
21	0	0	0	0	1/2	1/2	0
22	0	0	0	0	0	0	1

Matrices describing transition probabilities between different integron genotypes due to excision only ($\mathbf{M}^{exc}$) or excision followed by re-integration in the first position of the integron ($\mathbf{M}^{int}$), for a maximum integron length of two gene cassettes. Rows give the original genotype whereas columns give the genotype following integrase action. For example, the notation "21" refers to a genotype with cassette 2 adjacent to the promoter, followed by cassette 1. Zeros denote the absence of a cassette (only occurring at the end of an array). Note that integrons containing cassette duplications (11 and 22) can only emerge from integrons without such duplications in model extension 2 through HGT.

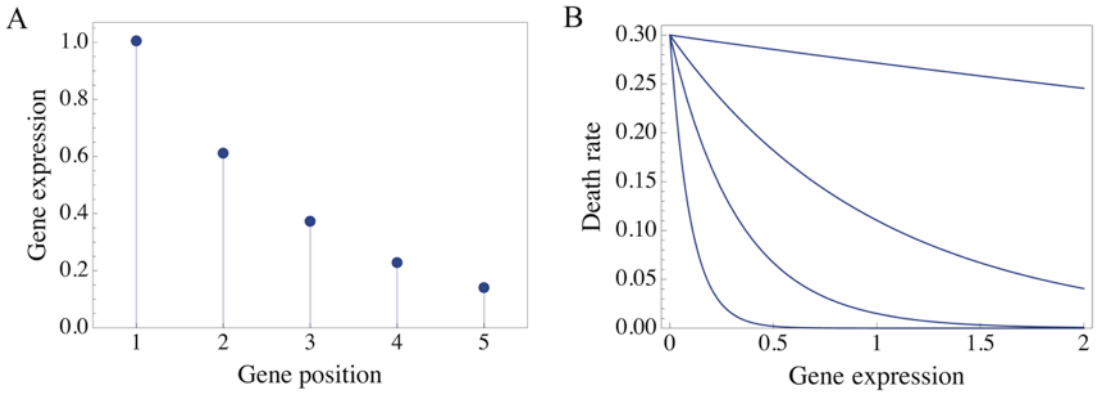

Figure S1. Illustration of the relation between gene position, gene expression and fitness in the model. Panel A shows how gene expression declines exponentially with increasing distance of a cassette gene from the promoter, with parameter $\beta = 0.5$. Plot B shows how total expression level of a gene providing resistance to a particular stressor affects the death rate caused by that stressor. The maximum expression level of a single-copy cassette is one, but expression levels higher than one can be achieved in Model Extension 2 when the integron contains duplicates of the same cassette. Parameters take the values $\eta_S = 0.3$ (maximum stress-induced death rate) and, from left to right, $\gamma = 10, 3, 1,$ and 0.1.

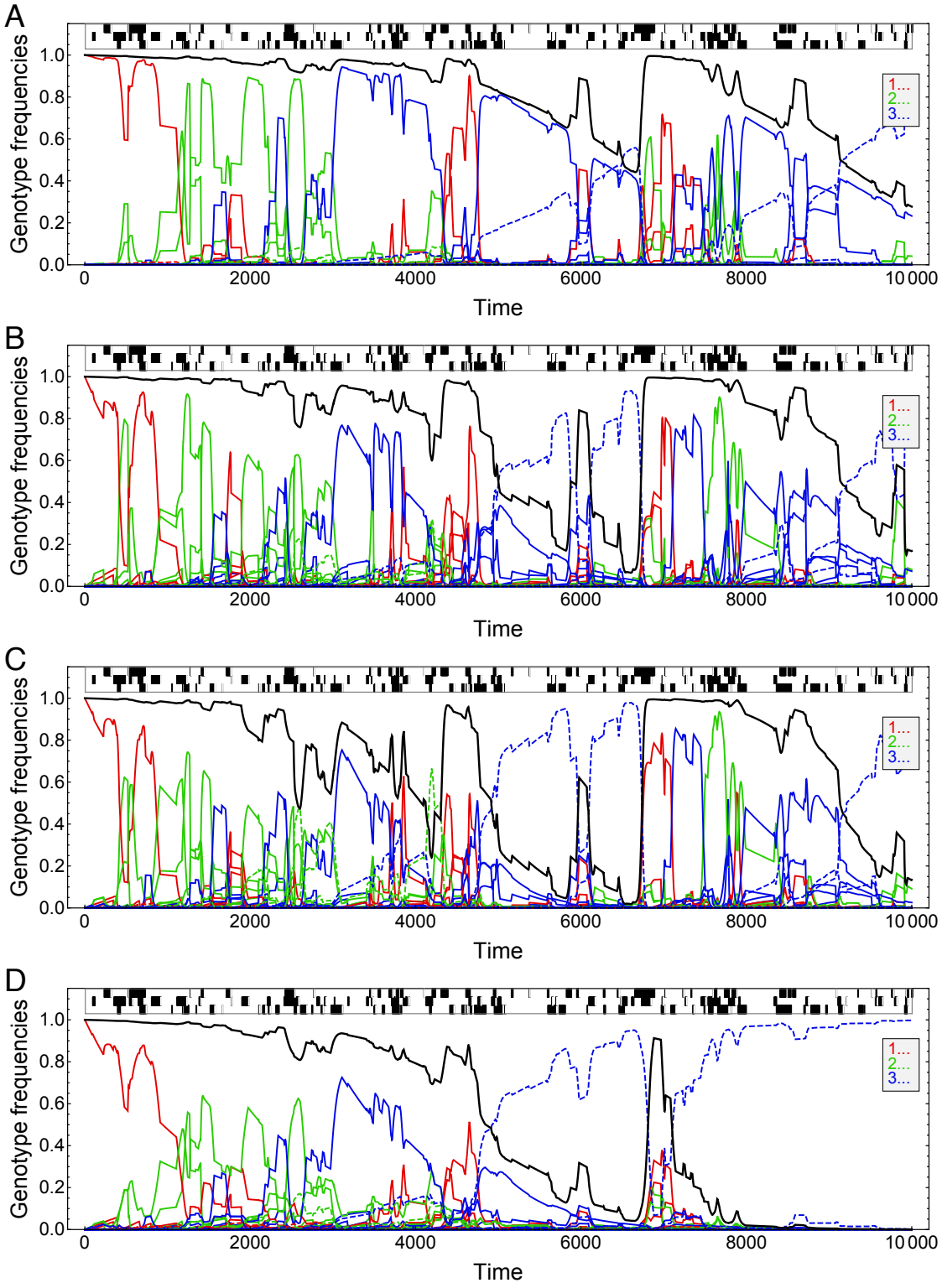

Figure S2. Example evolutionary dynamics with the basic model. The frequencies of the various genotypes are shown as in Figure 2A and to enable easy comparison, all simulations were run with the same randomly generated stress environment as in Figure 2. All parameters take standard values except (A) $\rho = 10^{-4}$, (B) $\theta = 0.2$, (C) $\eta_{\mathrm{I}} = 0.002$, and (D) $\eta_{\mathrm{S}} = 0.15$.

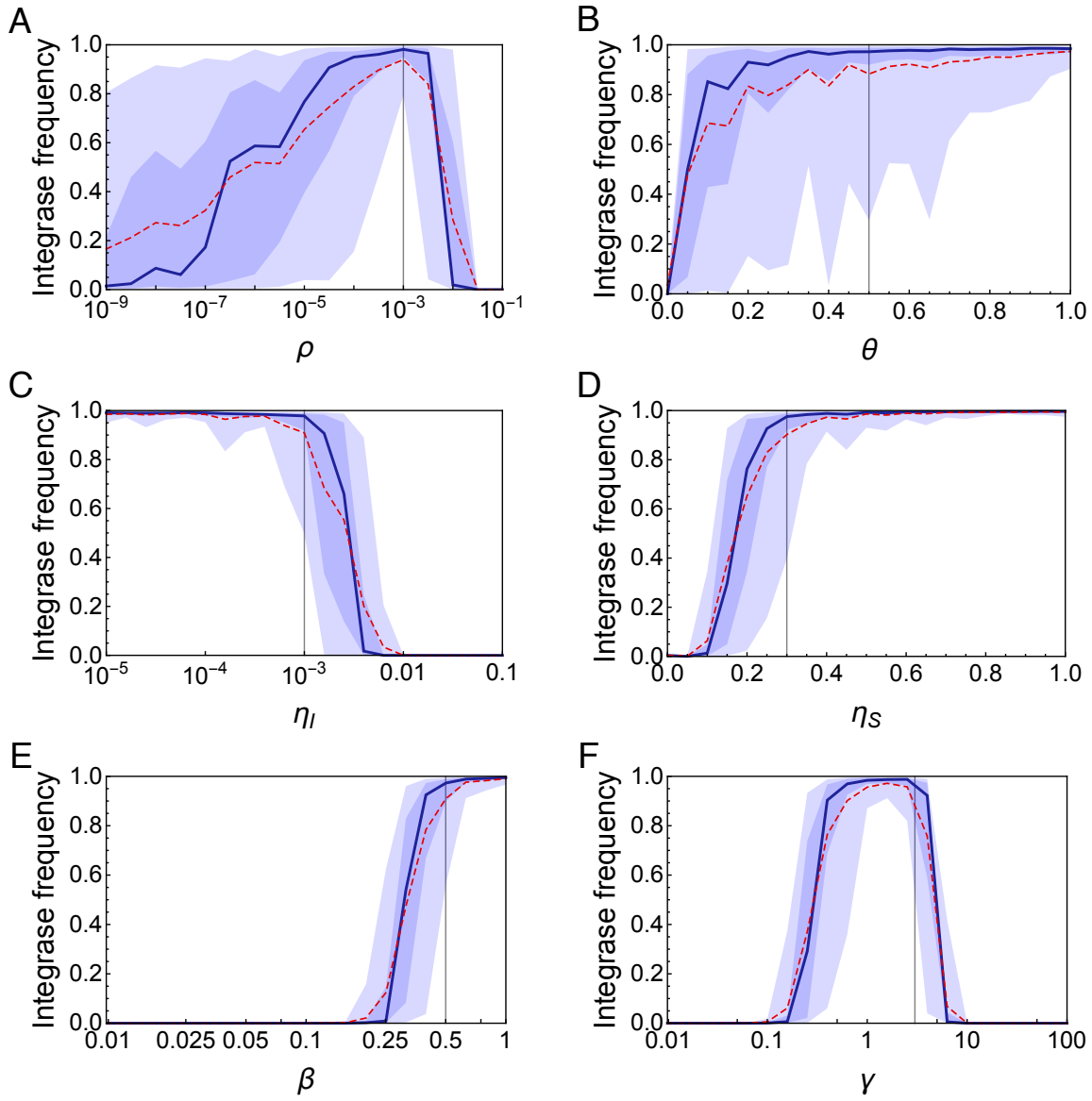

Figure S3. Impact of back-mutation from a non-functional to a functional integrase. The same simulations as for Figure 3 were run, except that a small rate of back-mutation was incorporated into the basic model. The rate of back-mutation was assumed to be $\mu_B = 10^{-8} = \mu/1000$.

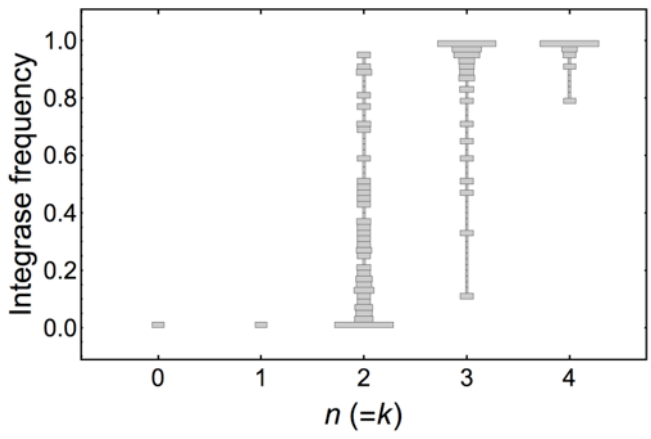

Figure S4. Impact of the number *n* of different gene cassettes on the selective maintenance of a functional integrase. Histograms of the final integrase frequency after 10,000 time units over 200 replicate simulations for different values of *n*. Note that *n* is equal to the number of stressors affecting the population and to the maximum length *k* of integrons. All parameters take standard values.

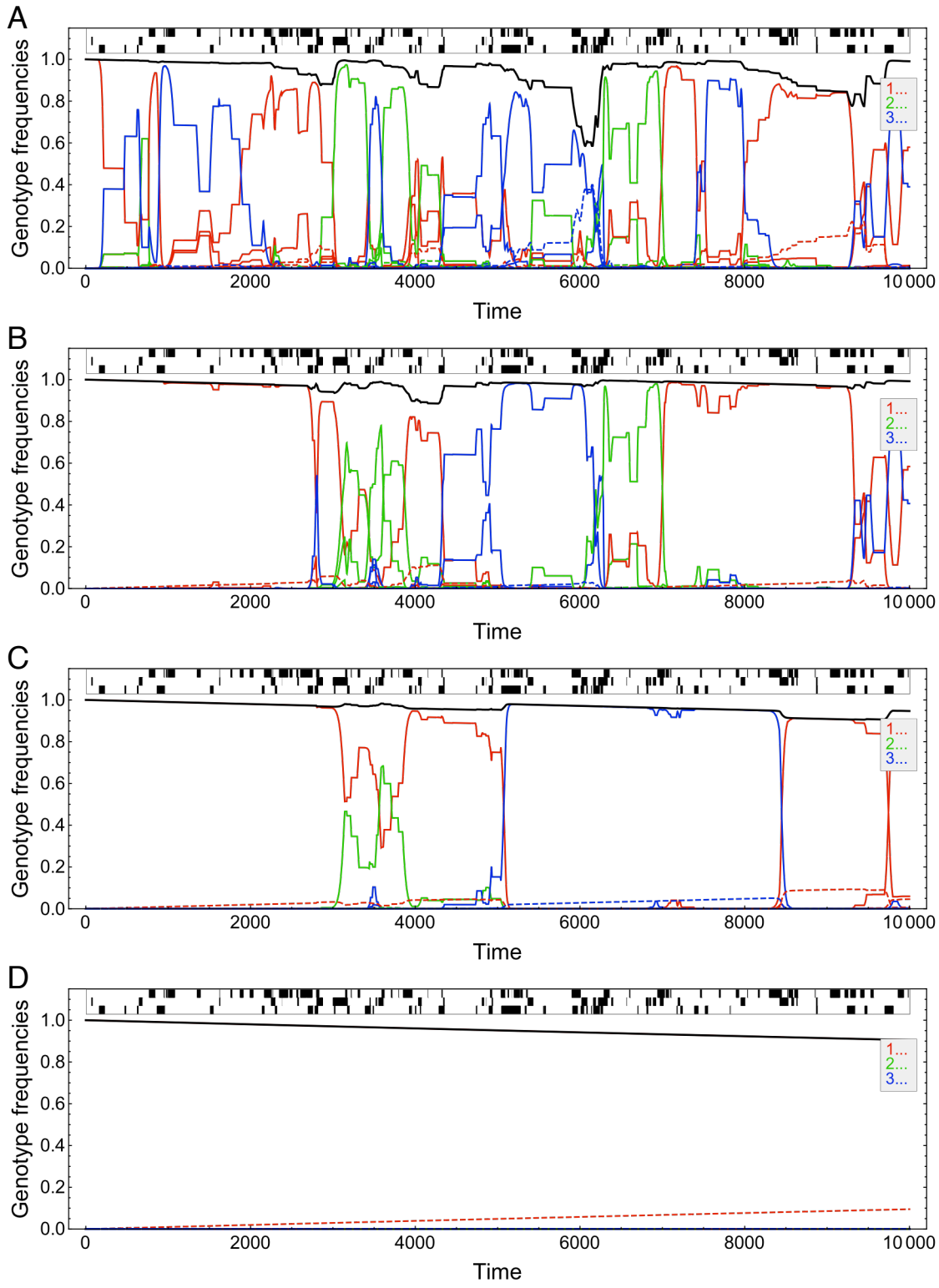

Figure S5. Example evolutionary dynamics with Model Extension 1 assuming stress-induced integrase expression. The frequencies of the various genotypes are shown as in Figures 2A and S2, and to enable easy comparison all simulations were run with the same randomly generated stress environment as in these previous figures. All parameters take standard values except (A) $\varphi = 0.05$, (B) $\varphi = 0.1$, (C) $\varphi = 0.15$, and (D) $\varphi = 0.2$.

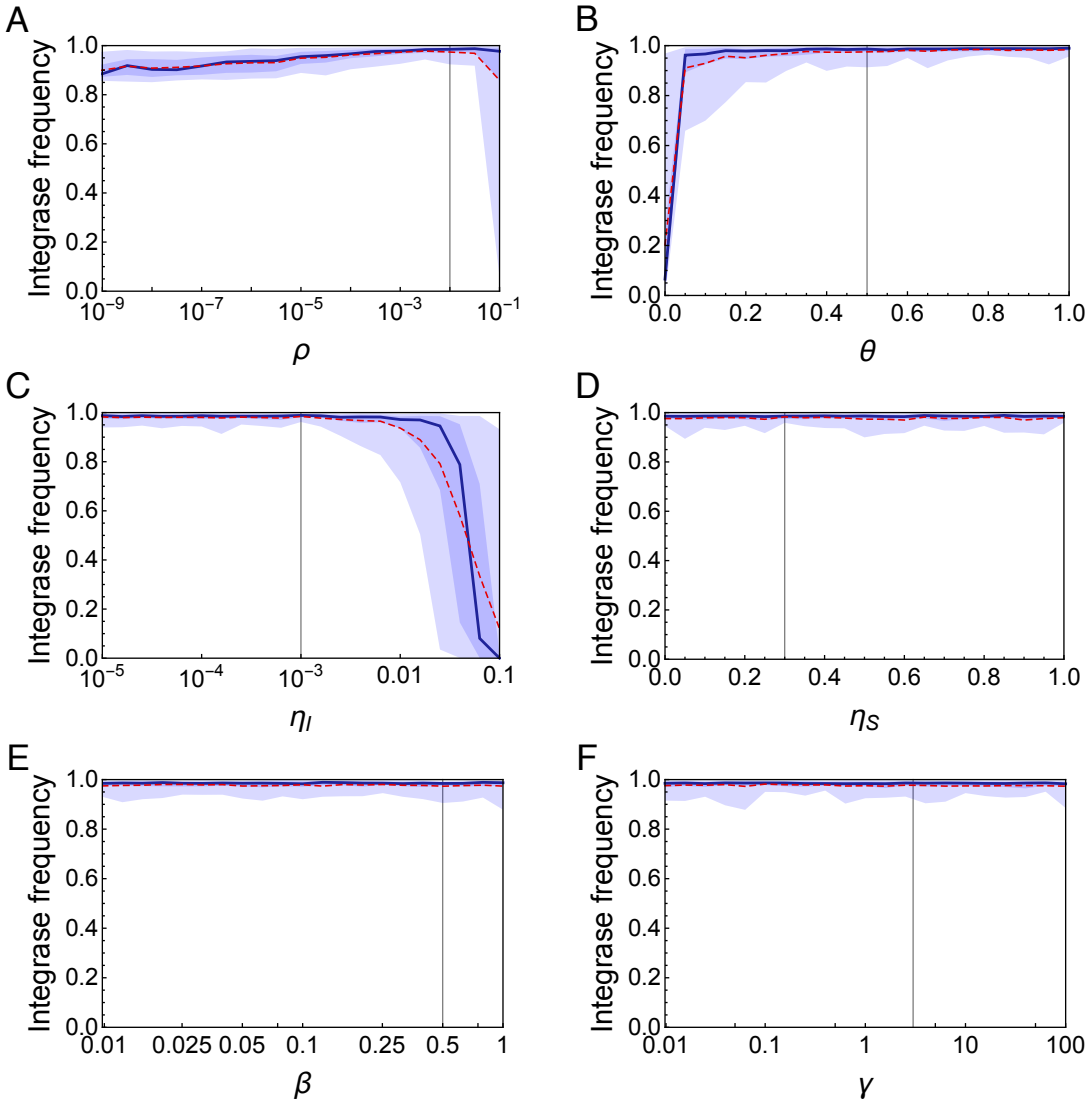

Figure S6. Impact of different parameters on the integrase frequency in Model Extension 1. Integrase transcription is stress-dependent, with threshold parameter $\varphi = 0.1$. The blue line gives the median frequency, the dashed red lines the average frequency and the shaded areas the interquartile and 90% interquantile range of the final integrase gene frequencies. All parameters take standard values except for the parameter varied. Grey vertical lines indicate the standard value of each parameter.

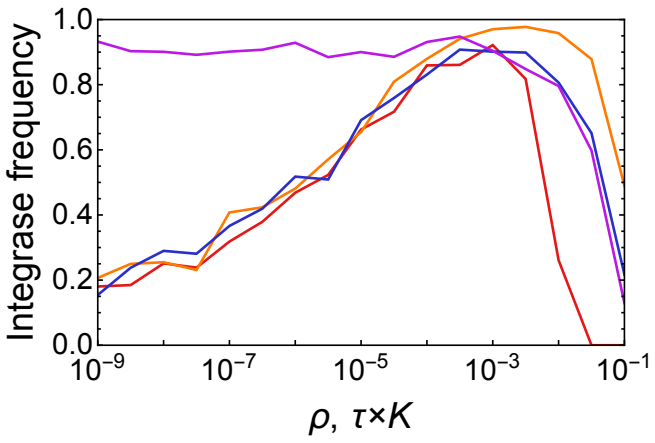

Figure S7. Comparison of integrase frequencies in different scenarios with and without horizontal gene transfer in Model Extension 2. Lines show the mean of the final integrase frequency across 100 replicate simulations. Red and orange: rate of integrase-mediated gene shuffling ρ is varied with either $\theta = 0.5$ (red) or $\theta = 1$ (orange). Blue and purple: rate of HGT τ is varied from 10^{-18} to 10^{-10} (corresponding to values of $\tau \times K$ shown on x-axis), with $\rho = 0$ (blue) or $\rho = 0.001$ (purple). All other parameters take standard values.

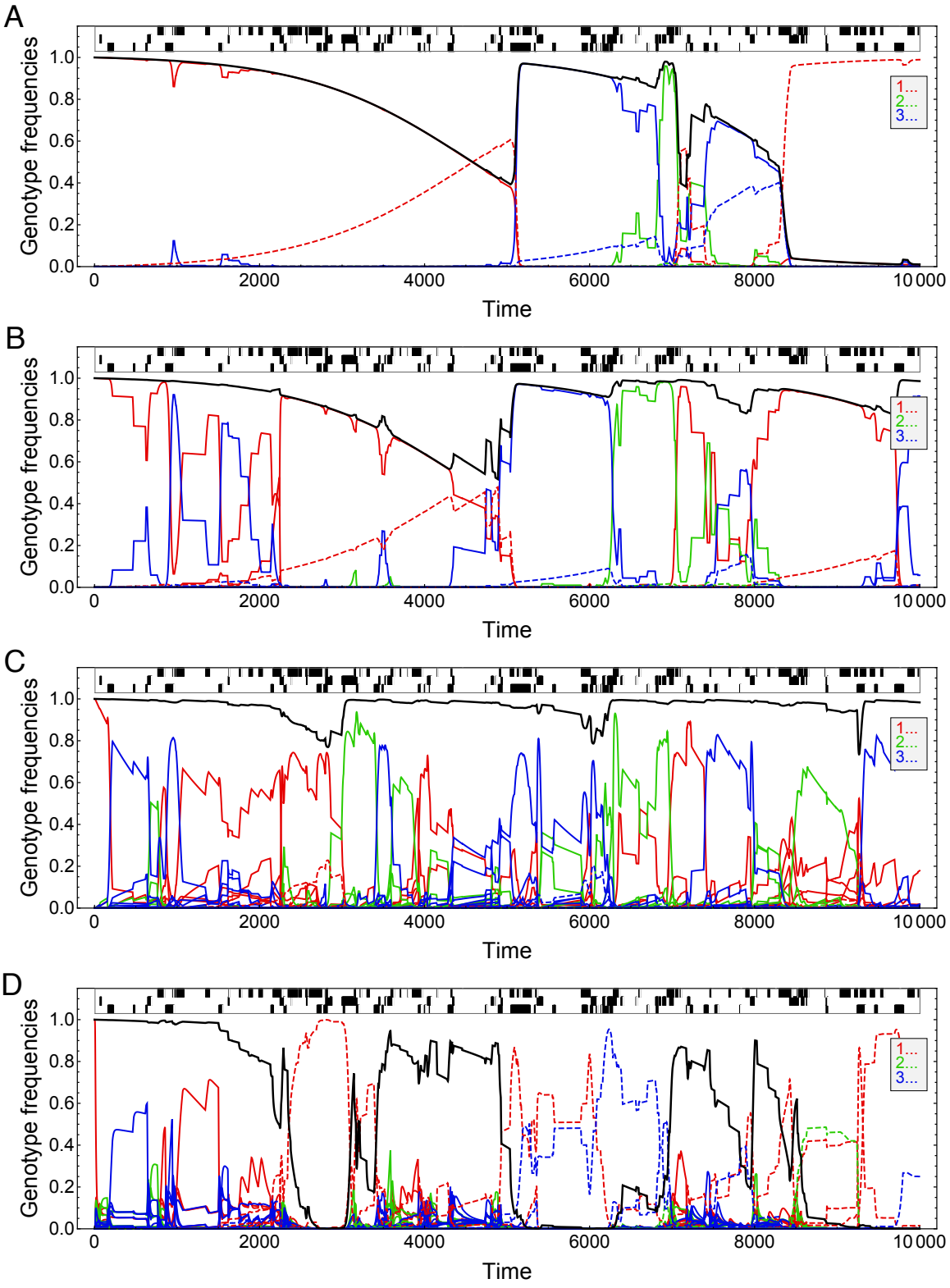

Figure S8. Example evolutionary dynamics with the Model Extension 2 assuming horizontal gene transfer. The frequencies of the various genotypes are shown as in Figures 2A, S2 and S5. To enable easy comparison all simulations were run with the same randomly generated stress environment as in these previous figures. All parameters take standard values except $\rho = 0$ and (A) $\tau = 10^{-16}$, (B) $\tau = 10^{-14}$, (C) $\tau = 10^{-12}$, and (D) $\tau = 10^{-10}$.