SUPPORTING INFORMATION

Disentangling the effects of selection and loss bias on gene dynamics

J. Iranzo, J.A. Cuesta, S. Manrubia, M.I. Katsnelson and E.V. Koonin

Contents

S1 The linear duplication-transfer-loss model with selection

Let us consider a population of cells whose genomes can host multiple copies of a given gene. In the most general setting of the model, a cell with k copies of the gene receives more copies via horizontal gene transfer (HGT) at rate h_k (h_0 is the HGT rate for cells that lack the gene). Furthermore, the gene copy number increases via duplications at rate d_k, and decreases (via deletion, gene loss, transposon excision, etc.) at rate l_k. The presence of the gene entails a fitness effect s_k for the cell, where positive values of s_k correspond to an increase in fitness. If the cell population is large, the dynamics of the gene copy number can be represented by the following system of differential equations, where n_k is the number of cells harboring k copies of the gene:

$$\begin{aligned} \frac{dn_0}{dt} &= (g - \Theta_0)\, n_0 + l_1 n_1 \\ \frac{dn_k}{dt} &= (g - \Theta_k)\, n_k + l_{k+1} n_{k+1} + (d_{k-1} + h_{k-1})\, n_{k-1} \end{aligned} \tag{S1}$$

where $\Theta_k = d_k + h_k + l_k - s_k$ and g is the basal growth rate of the cell population. The basal growth rate g affects the host population size but not the copy number distribution of the gene. Because we are interested in the latter, we make $g = 0$ for simplicity (note that the solution for the general case with $g \neq 0$ can be easily recovered by multiplying $n_k(t)$ by e^{gt}). We further simplify the model by assuming that each copy of the gene is an independent entity whose dynamics is not affected by the presence or absence of other copies within the genome, and therefore the total duplication and loss rates are linear with respect to k. Specifically, $d_k = kd$ and $l_k = kl$, where d and l are the duplication and loss rates per gene copy. In a similar way, we assume that the total fitness contribution linearly scales with the copy number, so that $s_k = ks$. The arrival of new copies via HGT is independent of the number of copies already present in the genome, and therefore $h_k = h$. Finally, we rescale all parameters by the loss rate l, which equals measuring time in units of expected loss events. After rescaling, the parameters of the model become $\rho = d/l$ for the duplication rate, $\beta = h/l$ for the HGT rate, and $\sigma = s/l$ for the fitness contribution, and the system (S1) becomes:

$$\begin{aligned} \frac{dn_0}{dt} &= -\beta n_0 + n_1 \\ \frac{dn_k}{dt} &= -(\alpha k + \beta)\, n_k + (k+1)\, n_{k+1} + \big(\rho\,(k-1) + \beta\big)\, n_{k-1} \end{aligned} \tag{S2}$$

where $\alpha = \rho - \sigma + 1$.

S2 Analytic solution of the model

S2.1 General solution

In order to find a solution of the model, let us introduce the generating function

$$G(z,t) \equiv \sum_{k=0}^{\infty} z^k n_k(t) \tag{S3}$$

Taking the partial derivative of the generating function with respect to t and substituting dn_k/dt from eqn. (S2) we obtain:

$$\frac{\partial G}{\partial t} = \frac{dn_0}{dt} + \sum_{k=1}^{\infty} z^k \frac{dn_k}{dt} = -\beta n_0 + n_1 + \sum_{k=1}^{\infty} z^k \left[-(\alpha k + \beta)\, n_k + (k+1)\, n_{k+1} + \big(\rho\,(k-1) + \beta\big)\, n_{k-1}\right] \tag{S4}$$

After some manipulation, this equation can be rewritten in terms of the generating function as

$$\frac{\partial G(z,t)}{\partial t} = (\rho z^2 - \alpha z + 1)\, \frac{\partial G(z,t)}{\partial z} + \beta(z-1) G(z,t) \tag{S5}$$

The equation of the characteristic curves of (S5) is

$$\frac{dz}{dt} = -\rho z^2 + \alpha z - 1 \tag{S6}$$

or equivalently

$$\frac{dz}{-\rho z^2 + \alpha z - 1} = dt \tag{S7}$$

In this equation, the denominator of the left member can be written as

$$\frac{1}{-\rho z^2 + \alpha z - 1} = \frac{1}{\rho a - a'} \left(\frac{1}{a - z} - \frac{\rho}{a' - \rho z}\right) \tag{S8}$$

where

$$a = \frac{\alpha + \sqrt{\alpha^2 - 4\rho}}{2\rho} \qquad \text{and} \qquad a' = \frac{\alpha - \sqrt{\alpha^2 - 4\rho}}{2} = a^{-1} \tag{S9}$$

Note that a and a' are real and positive as long as $\sigma < (1 - \sqrt{\rho})^2$, which defines the range of validity of the solution. In practice, this holds for all genetic parasites and most beneficial genes provided that the duplication rate is significantly smaller than the loss rate. For a neutral gene or in the absence of selection ($\sigma = 0$) it follows that $\rho a = \frac{1+\rho+|1-\rho|}{2} = \max\{1, \rho\}$ and $a' = \frac{1+\rho-|1-\rho|}{2} = \min\{1, \rho\}$. Combining (S7) and (S8), the characteristic curves can be written as

$$\frac{a' - \rho z}{a - z} = \xi\, e^{(\rho a - a')t} \tag{S10}$$

and solving for z,

$$z = \frac{a' - a\,\xi\, e^{(\rho a - a')t}}{\rho - \xi\, e^{(\rho a - a')t}} \tag{S11}$$

where ξ is a constant. On the characteristic curves, the differential equation (S5) becomes

$$\frac{dG(t)}{dt} = \beta\,(z-1)\, G(t) = \beta \left(\frac{a' - a\,\xi\, e^{(\rho a - a')t}}{\rho - \xi\, e^{(\rho a - a')t}} - 1\right) G(t) \tag{S12}$$

Equation (S12) admits a solution of the form

$$G(t) = F(\xi)\, W(t) \tag{S13}$$

with $F(\xi)$ an arbitrary function of ξ and $W(t)$ equal to

$$W(t) = \exp\left\{\beta \int \left(\frac{a' - a\,\xi\, e^{(\rho a - a')t}}{\rho - \xi\, e^{(\rho a - a')t}} - 1\right) dt\right\} \tag{S14}$$

A closed expression for $W(t)$ is obtained by solving the integral in eqn. (S14):

$$W(t) = e^{(a-1)\beta t}\, \left(\rho e^{-(\rho a - a')t} - \xi\right)^{\beta/\rho} \tag{S15}$$

For $t = 0$ we have $G(z,0) = F(\xi)\, W(0)$, from where it follows that $F(\xi) = \frac{G(z,0)}{W(0)}$. Moreover $z = \frac{a' - a\xi}{\rho - \xi}$, according to eqn. (S11). Substituting that into eqn. (S13) it results that

$$G(z,t) = G\left(\frac{a' - a\xi(z,t)}{\rho - \xi(z,t)}, 0\right)\, \left(\frac{\rho e^{-(\rho a - a')t} - \xi(z,t)}{\rho - \xi(z,t)}\right)^{\beta/\rho} e^{(a-1)\beta t} \tag{S16}$$

Replacing $\xi(z,t)$ by its expression in terms of z given by eqn. (S10) we finally get

$$G(z,t) = G\left(\frac{R(t) + z[1 - (\rho a + a')R(t)]}{1 - z\rho R(t)}, 0\right)\, \left(\frac{1 - a\rho R(t)}{1 - z\rho R(t)}\right)^{\beta/\rho} e^{(a-1)\beta t} \tag{S17}$$

where

$$R(t) = \frac{1 - e^{-(\rho a - a')t}}{\rho a - a' e^{-(\rho a - a')t}} \tag{S18}$$

and the particular choice for the term $G\left(\frac{R(t) + z[1 - (\rho a + a')R(t)]}{1 - z\rho R(t)}, 0\right)$ depends on the initial condition.

S2.2 Solution when the gene is initially absent

In the case of an initial population of size N with no copies of the gene ($n_0 = N$, $n_{k>0} = 0$), we have that $G(z,0) = N$ and eqn. (S17) becomes

$$G(z,t) = N\left(\frac{1 - a\rho R(t)}{1 - z\rho R(t)}\right)^{\beta/\rho} e^{(a-1)\beta t} \tag{S19}$$

S2.3 Solution when there are initially K copies per cell

If all cells harbor K copies of the gene ($n_K = N$, $n_{k\neq K} = 0$), the initial condition $G(z,0) = Nz^K$ applied to eqn. (S17) leads to

$$G(z,t) = N\left(\frac{R(t) + z[1 - (\rho a + a')R(t)]}{1 - z\rho R(t)}\right)^{K} \left(\frac{1 - a\rho R(t)}{1 - z\rho R(t)}\right)^{\beta/\rho} e^{(a-1)\beta t} \tag{S20}$$

In the asymptotic limit $t \to \infty$ the function $R(t) \to (\rho a)^{-1}$ and the generating function coincides, except for a multiplicative constant, with the case of an initial population free of parasites.

S3 Distribution of the gene copy number

S3.1 Generating function

In order to study the distribution of the gene copy number, it is useful to define the function

$$H(z,t) \equiv \frac{G(z,t)}{G(1,t)} \tag{S21}$$

The function $H(z,t)$ is the generating function associated to the gene copy number distribution:

$$H(z,t) = \frac{G(z,t)}{G(1,t)} = \frac{\sum_{k=0}^{\infty} z^k n_k(t)}{\sum_{k=0}^{\infty} n_k(t)} = \sum_{k=0}^{\infty} z^k \frac{n_k(t)}{\sum_{i=0}^{\infty} n_i(t)} = \sum_{k=0}^{\infty} z^k p_k(t) \tag{S22}$$

where $p_k(t) = \frac{n_k(t)}{\sum_{i=0}^{\infty} n_i(t)}$ is the fraction of cells in the population that have k copies of the gene at time t. If the gene is initially absent from the population, the generating function for the copy number distribution is obtained from eqn. (S19)

$$H(z,t) = \left(\frac{1-\rho R(t)}{1-z\rho R(t)} \right)^{\beta/\rho} \tag{S23}$$

S3.2 Copy number distribution: explicit expression

The copy number distribution is recovered from its generating function by expanding eqn. (S23) as a series of z and extracting the coefficients.

$$H(z,t) = \big(1-\rho R(t)\big)^{\beta/\rho} \left(\frac{1}{1-\rho R(t) z} \right)^{\beta/\rho} = \sum_{k=0}^{\infty} \big(1-\rho R(t)\big)^{\beta/\rho} C_k(\beta/\rho) \big(\rho R(t)\big)^k z^k \tag{S24}$$

where

$$C_k(x) = \frac{\prod_{j=0}^{k-1}(x+j)}{k!} = \frac{1}{k!} \frac{\Gamma(k+x)}{\Gamma(x)} \tag{S25}$$

Therefore, the fraction of cells with k copies of the parasite, p_k, is given by

$$p_k(t) = \big(1-\rho R(t)\big)^{\beta/\rho} \frac{\big(\rho R(t)\big)^k}{k!} \frac{\Gamma(k+\beta/\rho)}{\Gamma(\beta/\rho)} \tag{S26}$$

The limit $t \to \infty$ provides the asymptotic copy number distribution:

$$\lim_{t\to\infty} p_k(t) = \left(\frac{a-1}{a} \right)^{\beta/\rho} \frac{a^{-k}}{k!} \frac{\Gamma(k+\beta/\rho)}{\Gamma(\beta/\rho)} \tag{S27}$$

which is valid in the range $\sigma < (1-\sqrt{\rho})^2$, $\rho > 0$.

A case of particular interest corresponds to the absence of selection $\sigma = 0$. In such a neutral scenario, the copy number converges if $\rho < 1$ and adopts the following asymptotic distribution:

$$\lim_{\substack{\sigma\to 0 \\ t\to\infty}} p_k(t) = (1-\rho)^{\beta/\rho} \frac{\rho^k}{k!} \frac{\Gamma(k+\beta/\rho)}{\Gamma(\beta/\rho)} \tag{S28}$$

Genetic elements with a strict non-proliferative dynamics ($\rho = 0$) represent a special case in which the copy number follows a Poisson distribution with mean $\frac{\beta}{(1-\sigma)}$. In the limit $\rho \to 0$, the asymptotic expression of the generating function becomes:

$$\lim_{\substack{\rho\to 0 \\ t\to\infty}} H(z,t) = e^{\frac{\beta(z-1)}{1-\sigma}} \tag{S29}$$

The copy number distribution is obtained from the coefficients of the power series expansion of the exponential function:

$$\lim_{\substack{\rho\to 0\\ t\to\infty}} p_k(t) = e^{-\frac{\beta}{1-\sigma}} \frac{1}{k!} \left(\frac{\beta}{1-\sigma}\right)^k \tag{S30}$$

This expression is valid provided that $\sigma < 1$.

S3.3 Average copy number

The average copy number is obtained from the generating function as

$$\langle k(t)\rangle \equiv \sum_{k=0}^{\infty} k p_k(t) = \left.\frac{\partial H(z,t)}{\partial z}\right|_{z=1} \tag{S31}$$

which for the expression of $H(z,t)$ given by eqn. (S23) becomes

$$\langle k(t)\rangle = H(z,t)\left.\frac{\beta\, R(t)}{1-z\rho R(t)}\right|_{z=1} = \frac{\beta\, R(t)}{1-\rho R(t)} \tag{S32}$$

Substituting $R(t)$ by its expression in (S18) we obtain

$$\langle k(t)\rangle = \frac{\beta\left(1-e^{-(\rho a-a')t}\right)}{\rho(a-1)+(\rho-a')\,e^{-(\rho a-a')t}} \tag{S33}$$

S3.3.1 Neutral genes ($\sigma = 0$)

In a neutral scenario, $\sigma = 0 \Rightarrow a = \max\{1, 1/\rho\}, a' = \min\{1,\rho\}$ and the mean copy number becomes

$$\langle k(t)\rangle = \frac{\beta}{\rho-1}\left(e^{(\rho-1)t}-1\right) \tag{S34}$$

Three asymptotic regimes are possible depending on the value of ρ:

- If $\rho > 1$ (duplication bias), then $\langle k\rangle \sim \frac{\beta}{\rho-1}e^{(\rho-1)t}$, and the gene copy number explodes.
- If $\rho = 1$ (unbiased scenario), then taking the limit $\rho \to 1$ we get $\langle k\rangle = \beta t$, and the gene copies accumulate in the genome at a constant rate.
- If $\rho < 1$ (loss bias), them $\langle k\rangle \to \frac{\beta}{1-\rho}$, and the gene reaches a stable abundance in which losses are compensated by transfer of new copies.

S3.3.2 Deleterious genes ($\sigma < 0$)

In this case it always holds that $\rho a - a' > 0$ and $a > 1$. Regardless of the value of ρ, selection prevents unlimited proliferation of deleterious genes. The asymptotic value of the mean copy number is

$$\langle k\rangle = \frac{\beta}{\rho(a-1)} \tag{S35}$$

This expression is still valid if $\rho = 0$, and it takes the value $\langle k\rangle = \frac{\beta}{1-\sigma}$.

S3.3.3 Beneficial genes ($\sigma > 0$)

The solution of the model can be applied to beneficial genes provided that $0 < \sigma < (1 - \sqrt{\rho})^2$. In such case, the copy number reaches a stable asymptotic value if and only if $\rho < 1$, i.e. if there is a bias towards losses. The mean copy number is given by eqn. (S35), that is, it follows the same expression as in the case of deleterious genes.

S3.4 Fraction of genomes without the gene

The fraction of genomes that do not have any copy of the gene is obtained from the generating function as

$$p_0(t) = H(0,t) = (1 - \rho R(t))^{\beta/\rho} \tag{S36}$$

In the asymptotic limit $t \to \infty$

$$\lim_{t\to\infty} p_0(t) = \begin{cases} \left(\frac{a-1}{a}\right)^{\beta/\rho} & \text{if } \sigma < 0 \text{ or } \rho < 1 \\ 0 & \text{otherwise} \end{cases} \tag{S37}$$

In the particular case of a neutral gene in a loss biased scenario ($\sigma = 0$, $\rho < 1$), the asymptotic fraction of genomes without the gene is $p_0 = (1 - \rho)^{\beta/\rho}$.

S3.5 Characteristic time to equilibrium

Here we investigate how long it takes for a gene to reach the mean copy number described above after a perturbation. In the neutral scenario with loss bias ($\sigma = 0$, $\rho < 1$), the time-dependent term for the mean copy number decays exponentially as $e^{-(1-\rho)t}$ (eqn. S34). Therefore, the characteristic time for the relaxation to the equilibrium is $\tau = (1 - \rho)^{-1}$, where τ is measured in units of l^{-1}.

The case with selection is less simple due to the presence of the exponential term in the denominator of eqn. (S33). The equibrium is only approximately approached through an exponential decay if $\left|\frac{\rho - a'}{\rho(a-1)}\right| \ll 1$. Nevertheless, it is still possible to define a characteristic time as $\tau = (\rho a - a')^{-1}$, where, again, the time τ is measured in units of l^{-1}.

S3.6 Multiple gene families with equivalent biological properties

Let a and b be two gene families and let H_a and H_b be the generating functions for the copy number distribution of each family. The fraction of genomes with k copies of gene a is denoted by $p_{k,a}$, whereas the fraction of genomes with k' copies of gene b is denoted by $p_{b,k'}$. Because the linearity of the duplication-transfer-loss model, the copy numbers of a and b are independent random variables. Accordingly, the distribution of the sum copy number, $p_{a+b,k}$, is given by the convolution of the copy number distributions $p_{a,k}$ and $p_{b,k'}$. In terms of the generating function, the convolution of two distributions is given by the product of their generatig functions:

$$\begin{aligned} H_a(z,t)H_b(z,t) &= \left(\sum_{k=0}^{\infty} z^k p_{a,k}(t)\right)\left(\sum_{k'=0}^{\infty} z^{k'} p_{b,k'}(t)\right) = \\ &= \sum_{m=0}^{\infty} z^m \left(\sum_{k=0}^{m} p_{a,k}(t)\, p_{b,m-k}(t)\right) = \sum_{m=0}^{\infty} z^m\, p_{a+b,m}(t) = H_{a+b}(z,t) \end{aligned} \tag{S38}$$

The generating function H_{a+b} describes the fraction of genomes with a total of m copies, $p_{a+b,m}$, adding those from both gene families.

Let us now consider M families with the same values of σ and ρ (and therefore a and a'). The transfer rate β does not need to be the same for all families; there are instead M possibly different values β_i (with $i = 1, \ldots, M$) one for each gene family. The generating function for the total copy number, H_T, is obtained from eqn. (S23) as

$$H_T(z,t) = \prod_{i=1}^{M} H_i(z,t) = \left(\frac{1-\rho R(t)}{1-z\rho R(t)}\right)^{\sum_{i=1}^{M}\beta_i/\rho} \tag{S39}$$

In consequence, all the results presented above remain valid if we deal with a pool of similar gene families. The expressions for the pool of genes are obtained by making $\beta = \sum_{i=1}^{M}\beta_i$, that is, by inferring the total transfer rate as the sum of the transfer rates of each individual gene family. A practical implication is that the model parameters estimated from genomic data and the conclusions extracted from those estimates do not depend on the sequence similarity thresholds used to define gene families, as long as the members of the same family have similar selection to loss and duplication to loss ratios.

S3.7 Copy number distribution when the initial population has K copies per cell

Following section S2.3, we present here the generating function that describes the transient dynamics of the copy number distribution when all cells in the population initially carry K copies of the gene. Using equations (S20) and (S21), the generating function for the copy number distribution becomes

$$H(z,t) = \left(\frac{R(t) + z[1-(\rho a + a')R(t)]}{1+(1-\rho a - a')R(t)}\right)^{K} \left(\frac{1-\rho R(t)}{1-z\rho R(t)}\right)^{K+\beta/\rho} \tag{S40}$$

S4 Equivalence between neutrality and selection

S4.1 Equivalence of the asymptotic distributions

Upon normalization by the loss rate, the duplication-transfer-loss model with selection is characterized by three parameters: the duplication to loss ratio ρ, the HGT to loss ratio β, and the selection to loss ratio σ. To simplify notation, we defined the composite parameter a, which absorbs the effects of selection. In the stationary state, the copy number distribution is described by eqn. (S27), which depends on ρ, β, and a. It is easy to see, however, that the same copy number distribution can be expressed as a function of only two parameters:

$$p_k = (1-\phi)^{\psi} \frac{\phi^k}{k!} \frac{\Gamma(k+\psi)}{\Gamma(\psi)} \tag{S41}$$

where we defined

$$\begin{aligned} \psi &= \beta/\rho \\ \phi &= 1/a \end{aligned} \tag{S42}$$

There are infinite combinations of the model parameters ρ, σ and β that result in the same ψ and ϕ, and any of those combinations will generate the same copy number distribution in the stationary state. Such degree of freedom makes it impossible to determine the model parameters only from an observed distribution, without prior knowledge of at least one of the parameters. Conversely, it implies that genes with different fitness effects, duplication rates and HGT rates may exhibit identical copy number distributions.

The equivalence of copy number distributions also affects the distinction between neutral genes and genes whose copy number is under selection (henceforth denoted as "selected" genes). To illustrate this point, let us rename the parameters of the model in the absence of selection ($\sigma = 0$) as ρ_{eq} and β_{eq}. Comparing the expressions in (S28) and (S41) it is clear that the distributions of selected and neutral genes are identical, as long as $\psi = \beta_{eq}/\rho_{eq}$ and $\phi = \rho_{eq}$. Therefore, there is an equivalence between the asymptotic distributions of neutral and selected elements, that in terms of the model parameters is expressed by

$$\begin{aligned} \beta/\rho &= \beta_{eq}/\rho_{eq} = \psi \\ 1/a &= \rho_{eq} = \phi \end{aligned} \tag{S43}$$

Extracting σ from a, the second expression in (S43) can be written as

$$\sigma = \frac{(1-\rho_{eq})(\rho_{eq}-\rho)}{\rho_{eq}} \tag{S44}$$

Given a set of parameters estimated from genomic parasite distributions (typically under the assumption of neutrality), the expressions (S43-S44) can be used to re-estimate the fitness cost, proliferation rate and transfer rate under a general scenario with selection (or vice versa), provided that there is prior knowledge on the value of at least one of those parameters.

The equivalence between the stationary copy number distributions of neutral and selected parasites also holds in the special case $\rho = 0$ described by eqn. (S30). In a strict non-proliferative scenario, the stationary copy number distribution can be described by a single parameter $\varphi = \frac{\beta}{(1-\sigma)} = \beta_{eq}$. Any combination of β, σ and β_{eq} that yields the same value of φ produces the same stationary distribution. Importantly, distributions of neutral genes with $\rho_{eq} = 0$ can only be matched to distributions of selected genes with $\rho = 0$ and vice versa. As a result, the cases $\rho = 0$ and $\rho > 0$ are mutually distinguishable regardless of the presence or lack of selection.

S4.2 Equivalence of transient distributions

Let us consider the transient distribution given by eqn. (S26). Now let us suppose that we have access to a population for which we do not have any knowledge on the time variable t. The copy number distribution for such population will have the form:

$$p_k = (1-\Phi)^{\psi}\,\frac{\Phi^k}{k!}\,\frac{\Gamma(k+\psi)}{\Gamma(\psi)} \tag{S45}$$

where $\psi = \beta/\rho$ as before and $\Phi = \rho R(t)$. In general, the value of Φ grows from $\Phi = 0$ when $t = 0$ to $\Phi = a^{-1}$ when $t \to \infty$. Comparing eqn. (S41) and (S45) it is clear that the family of distributions that describe the transient states is the same as the family of stationary distributions. As a result, it is impossible to say if a population has reached the stationary state just by looking at a single snapshot of the copy number distribution.

S4.3 Recovery time after perturbations

Beyond their equivalence in terms of the copy number distribution, the dynamics of neutral and selected genes have different characteristic times that may allow distinguishing them after a perturbation. Following Section S3.5, the characteristic time is $\tau_{eq} = (1-\rho_{eq})^{-1}$ for a neutral gene and $\tau = (\rho a - a')^{-1}$ for a gene under selection. In the case of two genes with the same stationary distribution we have

$$\frac{\tau}{\tau_{eq}} = \frac{1-\rho_{eq}}{\rho a - a'} = \frac{a-1}{a(\rho a - a')} = \frac{1-\rho-\sigma+\sqrt{(1+\rho-\sigma)^2-4\rho}}{\sqrt{(1+\rho-\sigma)^2-4\rho}\left(1+\rho-\sigma+\sqrt{(1+\rho-\sigma)^2-4\rho}\right)} \tag{S46}$$

which tends to 1 in the limit $\sigma \to 0$. The derivative of this expression with respect to σ has the form

$$\frac{d(\tau/\tau_{eq})}{d\sigma} = \frac{1-\rho-\sigma}{((1+\rho-\sigma)^2-4\rho)^{3/2}} \tag{S47}$$

For $\rho < 1$, the quotient τ/τ_{eq} monotonically increases with σ in the whole range of validity of the solution, with $\tau/\tau_{eq} = 1$ for $\sigma = 0$. As a result, deleterious genes reach their stationary state faster than their "equivalent" neutral genes, whereas beneficial genes are slower. For $\rho > 1$, only deleterious genes reach a stationary state. In that case the quotient τ/τ_{eq} has a minimum at $\rho = 1-\sigma$, where it takes the value $\tau/\tau_{eq} = \left(2-2\sigma+2\sqrt{\sigma(\sigma-1)}\right)^{-1} < 1/2$. Again, deleterious genes reach the stationary state faster than their equivalent neutral genes.

S5 Maximum likelihood estimation of the duplication to loss ratio

In this section we obtain a maximum likelihood estimate of the parameter ρ using the copy number of genes that are present in a single genome within an ATGC (henceforth called ORFans). It is reasonable to assume that those genes are the result of a single HGT event to the recipient genome and that multiple copies are the result of duplication events. Based on such assumptions, we first obtain the likelihood function for ρ given a collection of ORFans, their copy numbers and the time since the last branching event in the phylogenetic tree. Then we describe how the likelihood-based approach was implemented to estimate ρ from the genomic dataset.

S5.1 Likelihood function

To derive the expresion of the log-likelihood we need the probability $P(k_i|\rho, k_i > 0)$ that k_i copies of an element i are present in a genome, conditioned to the existence of at least one copy. Conditioning on the time when the element arrived in the genome we can write:

$$P(k_i|\rho, k_i > 0) = \int_0^\infty P(k_i|t, \rho, k_i > 0)\, P(t|\rho, k_i > 0)dt \tag{S48}$$

The first term in the integral is equal to

$$P(k_i|t, \rho, k_i > 0) = \frac{P(k_i|t, \rho)}{1 - P(0|t, \rho)} \tag{S49}$$

The second term in the integral is obtained applying the Bayes theorem to the probability that at least one copy of the element persists in the genome after a time t:

$$P(t|\rho, k > 0) = \frac{P(k_i > 0|t, \rho)\, P(t)}{\int_0^\infty P(k_i > 0|u, \rho)\, P(u)\, du} \tag{S50}$$

where the integral in the denominator is done with respect to the time variable u. Assuming that the element arrived after the last branching event in the phylogenetic tree, we take a non-informative prior $P(u) = 1/T_i$ in the interval $0 \leq u \leq T_i$, where T_i is the time since the last branching event. With this assumption, equation (S50) becomes

$$P(t|\rho, k > 0) = \frac{1 - P(0|t, \rho)}{\int_0^{T_i}(1 - P(0|u, \rho))\, du} \tag{S51}$$

The expression for $P(k_i|\rho, k_i > 0)$ results from combining equations (S48), (S49), and (S51)

$$P(k_i|\rho, k_i > 0) = \frac{\int_0^{T_i} P(k_i|t, \rho)\, dt}{\int_0^{T_i}(1 - P(0|t, \rho))\, dt} \tag{S52}$$

Because we assume that genes that are present in a single genome within an ATGC (ORFans) are the result of a single HGT event, the expression for $P(k_i|t, \rho)$ can be derived from a duplication-transfer-loss process with $\beta = 0$ and initial condition $p_k(0) = \delta_{1,k}$. A series expansion of equation (S40) with $\beta = 0$ provides, in the neutral limit $\sigma \to 0$, the result

$$P(0|t, \rho) = p_0(t) = R(t) \tag{S53}$$

$$P(k|t, \rho) = p_k(t) = (1 - R(t))\,(1 - \rho R(t))\,(\rho R(t))^{k-1} \tag{S54}$$

where

$$R(t) = \frac{1 - e^{-(1-\rho)t}}{1 - \rho e^{-(1-\rho)t}} \tag{S55}$$

Given a set of ORFans as the ones described above, $i \in \mathcal{I}$, with copy number k_i, and being T_i the length of the terminal branch leading to the host genome in the phylogenetic tree, the loglikelihood function for the paremeter ρ becomes

$$LL(\rho) = \sum_{i \in \mathcal{I}} \log \left[\int_0^{T_i} (1 - R(t)) \left(1 - \rho R(t)\right) \left(\rho R(t)\right)^{k_i - 1} \right] - \sum_{i \in \mathcal{I}} \log \left[\int_0^{T_i} (1 - R(t)) \right] \tag{S56}$$

S5.2 Practical implementation

Because the parameters of the model were normalized by the loss rate, the appropriate units for the branch lengths T_i in eqn. (S56) are those corresponding to the gene loss time scale. Accordingly, we used the branch lengths provided by the software COUNT, which are based on gene copy number divergence due to duplication, loss and transfer events. Moreover, we added a second variable ω to the likelihood function that accounts for the proportionality between the time scale at which genes are lost and the average time scale at which the gene copy number diverges in sister lineages. Specifically, if B_i is the branch length provided by COUNT, we made $T_i = \omega B_i$ and maximized the function

$$LL(\rho, \omega) = \sum_{i \in \mathcal{I}} \log \left[\int_0^{B_i \omega} (1 - R(t)) \left(1 - \rho R(t)\right) \left(\rho R(t)\right)^{k_i - 1} \right] - \sum_{i \in \mathcal{I}} \log \left[\int_0^{B_i \omega} (1 - R(t)) \right] \tag{S57}$$

The maximization of the loglikelihood function $LL(\rho, \omega)$ was carried out using the Nelder-Mead simplex method as implemented in MATLAB R2016b and yielded the estimates $\rho = 0.126$ and $\omega = 3.15$.

S6 Model with proliferation bursts

This section explores a modified model in which the duplication-transfer-loss dynamics is punctuated by bursts of proliferation that occur at exponentially distributed intervals with characteristic time $T = 1/\phi$ and reset the copy number to K. Specifically, we are interested in the mean copy number generated by such model in the long-term average, which is equivalent to the mean copy number expected for a pool of genomes with independent burst histories.

The mean copy number averaged in time is equal to

$$\langle\langle k\rangle\rangle = \int_0^\infty dt \sum_{k=0}^{\infty} \left(k\, p_k^{(K)}(t)\right)\, \phi\, e^{-\phi t} \tag{S58}$$

where $p_k^{(K)}(t)$ is the copy number distribution with initial condition $p_k^{(K)}(0) = \delta_{k,K}$ and the term $\phi e^{-\phi t}$ denotes the probability that the time interval since the most recent burst is equal to t. To simplify the calculation, we first obtain the sum term, which corresponds to the time-dependent mean copy number:

$$\langle k(t)\rangle = \sum_{k=0}^{\infty} k\, p_k^{(K)}(t) = \left.\frac{\partial H(z,t)}{\partial z}\right|_{z=1} \tag{S59}$$

The generating function $H(z,t)$ that corresponds to the initial condition $p_k^{(K)}(0) = \delta_{k,K}$ is given by eqn. (S40). After some manipulation, it results that the time-dependent mean copy number can be expressed as

$$\langle k(t)\rangle = \kappa_1(t) + K\,\kappa_2(t) \tag{S60}$$

where

$$\begin{aligned} \kappa_1(t) &= \frac{\beta R(t)}{1-\rho R(t)} \\ \kappa_2(t) &= \frac{1-(\rho a + a')R(t)}{1+(1-\rho a - a')R(t)} + \frac{\rho R(t)}{1-\rho R(t)} \end{aligned} \tag{S61}$$

Substituting this into eqn. (S58), the integral for the time-averaged mean copy number becomes:

$$\langle\langle k\rangle\rangle = \phi \int_0^\infty \kappa_1(t) e^{-\phi t} dt + K\phi \int_0^\infty \kappa_2(t) e^{-\phi t} dt \tag{S62}$$

This integral admits a convoluted solution in terms of hypergeometric functions. However, for practical purposes, its value must be calculated numerically. Note that if the characteristic interval between bursts is much larger than the characteristic time of the transient dynamics, $\phi << \rho a - a'$, it is possible to approximate this model with exponentially separated bursts by a similar model with regularly separated bursts. Such approach provides the following approximated value for the average mean copy number:

$$\langle\langle k\rangle\rangle \approx \frac{\beta}{\rho(a-1)} + \phi\, K\, \frac{\ln(a'/\rho)}{(a'-\rho)(a-1)} - \phi\, \frac{\beta}{a'-\rho} \ln\left(\frac{\rho(a-1)}{\rho a - a'}\right) \tag{S63}$$

S7 Correlations with effective population size

The effective population sizes (N_e) were inferred for each ATGC as described in (1). In short, non-synonymous to synonymous nucleotide substitution ratios (dN/dS) were evaluated for each ATGC using concatenated sequences of core genes. The values of dN/dS were subsequently translated to effective population sizes by applying the expression $\frac{dN}{dS} \approx \frac{N_e s_c}{1-e^{-N_e s_c}}$ (2), where the selection coefficient of core genes s_c was set such that the effective population size for Escherichia coli is 10^9.

The correlations between the model parameters and N_e follow the same pattern as those between the model parameters and genome size. Specifically, there is a significant association between N_e and h/l (Spearman's $\rho = 0.40$, $p = 0.017$) as well as with the number of ORFan families per genome (Spearman's $\rho = 0.64$, $p < 10^{-4}$). In contrast, N_e is not associated with d/l (Spearman's $\rho = 0.16$, $p = 0.35$) or with the fraction of ORFan families with more than one copy (Spearman's $\rho = 0.20$, $p = 0.26$).

To determine which variable, genome size or N_e, is responsible for the trends observed in h/l we performed a partial correlation analysis. When both variables are considered, the association between h/l and N_e disappears ($p = 0.81$), which implies that variations in genome size are the primary cause of the variations in h/l.

S8 Supplementary Figures

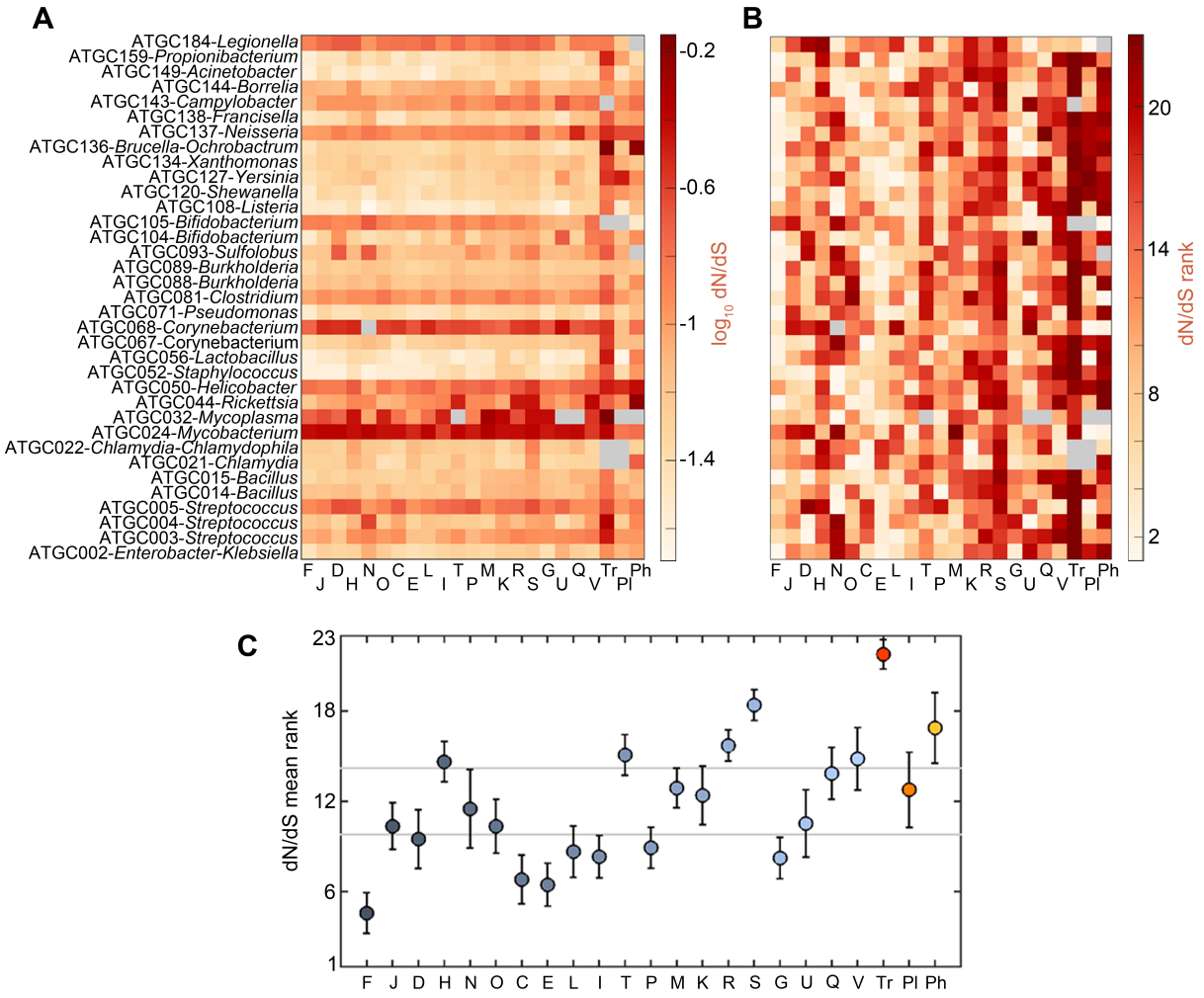

Figure S1. Non-synonymous to synonymous mutation ratio (dN/dS) for different gene categories. A: The color intensity represents, in logarithmic scale, the median dN/dS of the ATGC-COGs that belong to a specific ATGC and category. The cases where the number of informative pairs was too small to infer the dN/dS are indicated in grey. B: ATGC-wise rank of functional categories according to their dN/dS. Lower ranks correspond to lower dN/dS. C: Comparison of the dN/dS mean ranks among functional categories. Circles are the mean ranks, averaged across ATGCs, error bars represent the standard error of the mean. The horizontal grey lines indicate the theoretical 95% CI expected for the means of a null model where the dN/dS of all categories follow the same distribution (points above/below this interval indicate that the dN/dS of a category is significantly higher/lower than average).

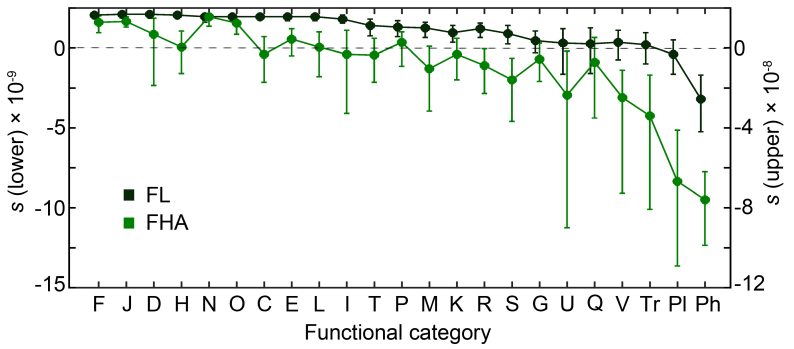

Figure S2. Selection coefficients in free-living (FL) and facultative host-associated (FHA) microbes, under the assumption that the intrinsic duplication to loss ratio (d/l) is the same in both lifestyles. The designations of functional classes in the x-axis are the same as in Figure 1 and Table 1. Error bars were obtained by combining the 95% CI for the median d/l_e and the intrinsic d/l. The scale on the left axis corresponds to a lower estimate using a loss rate $d = 5 \times 10^{-9}$ per gene per generation, the scale on the right axis corresponds to an upper estimate with $d = 4 \times 10^{-8}$ per gene per generation.

S9 Supplementary Tables

Table S1. Contributions of selection and the duplication/loss ratio to the evolution of different functional subcategories of genes within category N (secretion and motility).

			s ($\times 10^{-8}$)	
	d/l_e	s/l	Lower	Upper
N, secretion and motility	0.247	0.37	0.19	1.49
N(i), flagellum components	0.261	0.38	0.19	1.54
N(ii), cellulose production and glycosyltransferases	0.141	0.10	0.05	0.39

The table shows the estimated values of the effective duplication/loss ratio (d/l_e), selection to loss ratio (s/l) and selection coefficient (s) for different functional categories of genes The s/l values were calculated assuming an intrinsic duplication/loss ratio $d/l = 0.125$. Loss rates equal to 5×10^{-9} and 4×10^{-8} per gene per generation were used to obtain the lower and upper estimates of s, respectively.

References

1. Sela, I., Wolf Y.I. & Koonin, E.V. Theory of prokaryotic genome evolution. *Proc. Natl. Acad. Sci. USA*, **113**, 11399-11407 (2016).

2. Kryazhimskiy S. & Plotkin, J.B. The population genetics of dN/dS. *PLoS Genet.*, **4**, e1000304 (2008).