

Disentangling the effects of selection and loss bias on gene dynamics

Jaime Iranzo^a, José A. Cuesta^{b,c,d}, Susanna Manrubia^e, Mikhail I. Katsnelson^f, and Eugene V. Koonin^{a,1}

^aNational Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; ^bGrupo Interdisciplinar de Sistemas Complejos, Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), Spain; ^cInstitute for Biocomputation and Physics of Complex Systems, Universidad de Zaragoza, 50018 Zaragoza, Spain; ^dInstitute of Financial Big Data, Universidad Carlos III de Madrid-Banco de Santander, 28903 Getafe (Madrid), Spain; ^eGrupo Interdisciplinar de Sistemas Complejos, National Biotechnology Centre, Consejo Superior de Investigaciones Científicas, 28049 Madrid, Spain; and ^fInstitute for Molecules and Materials, Radboud University, Nijmegen 6525AJ, The Netherlands

Contributed by Eugene V. Koonin, June 1, 2017 (sent for review March 24, 2017; reviewed by Sergei Maslov and Dennis Vitkup)

We combine mathematical modeling of genome evolution with comparative analysis of prokaryotic genomes to estimate the relative contributions of selection and intrinsic loss bias to the evolution of different functional classes of genes and mobile genetic elements (MGE). An exact solution for the dynamics of gene family size was obtained under a linear duplication–transfer–loss model with selection. With the exception of genes involved in information processing, particularly translation, which are maintained by strong selection, the average selection coefficient for most nonparasitic genes is low albeit positive, compatible with observed positive correlation between genome size and effective population size. Free-living microbes evolve under stronger selection for gene retention than parasites. Different classes of MGE show a broad range of fitness effects, from the nearly neutral transposons to prophages, which are actively eliminated by selection. Genes involved in antiparasite defense, on average, incur a fitness cost to the host that is at least as high as the cost of plasmids. This cost is probably due to the adverse effects of autoimmunity and curtailment of horizontal gene transfer caused by the defense systems and selfish behavior of some of these systems, such as toxin–antitoxin and restriction modification modules. Transposons follow a biphasic dynamics, with bursts of gene proliferation followed by decay in the copy number that is quantitatively captured by the model. The horizontal gene transfer to loss ratio, but not duplication to loss ratio, correlates with genome size, potentially explaining increased abundance of neutral and costly elements in larger genomes.

mobile genetic elements | selection | gene loss | horizontal gene transfer | antiparasite defense

In the wake of the genomic revolution, quantitative understanding of the roles that ecological and genetic factors play in determining the size, composition, and architecture of genomes has become a central goal in biology (1–3). The vast number of prokaryotic genomes sequenced to date reveals a great diversity of sizes, which range from about 110 kb and 140 protein coding genes in the smallest intracellular symbionts (4) to almost 15 Mb and more than 10,000 genes in the largest myxobacteria (5). Beyond a core of ~100 nearly universal genes, the gene complements of bacteria and archaea are highly heterogeneous (6–8). Remarkably, 10–20% of the genes in most microbial genomes are ORFs, that is, genes that have no detectable homologs in other species and are replaced at extremely high rates in the course of microbial evolution (9, 10). Furthermore, all but the most reduced genomes host multiple and diverse parasitic genetic elements, such as transposons and prophages that collectively compose the so-called microbial mobilome (11).

The evolution of microbial genomes is generally interpreted in terms of the interplay between three factors: (i) gene gain, via horizontal gene transfer (HGT) and gene duplication; (ii) gene loss, via deletion; and (iii) natural selection that affects the fixation and maintenance of genes (8, 12). The intrinsic bias toward DNA deletion (and hence gene loss) that characterizes mutational processes in prokaryotes (as well as eukaryotes) results in nonadaptive ge-

nome reduction (13), whereas selection contributes to maintaining slightly beneficial genes (14). In agreement with this model, the strength of purifying selection, as measured by the ratio of non-synonymous to synonymous variation, positively correlates with the genome size (15, 16). However, when it comes to interpreting the genome composition, the picture is complicated by the fact that selection can also lead to adaptive genome reduction by removing pseudogenes (17), costly genetic parasites, and accessory genes, which are dispensable under stable environmental conditions (18, 19). Conversely, the increased propensity of some gene families to be horizontally transferred might suffice to ensure their persistence beyond the effects of selection and intrinsic loss bias (20). Rather than being minor deviations from a general trend, nonuniform levels of selection and horizontal gene transfer affecting different families and classes of genes appear to be essential to explain the abundance distributions and evolutionary persistence times of genes (10, 12). Accordingly, a quantitative assessment of the fitness costs and benefits for different classes of genes is essential to attain an adequate understanding of the evolutionary forces that shape genomes.

The magnitude and even the sign with which the presence (or absence) of a gene contributes to the fitness of an organism are not constant in time. For example, the metabolic cost incurred by the replication, transcription, and translation of a gene strongly depends on the cell growth rate and the gene expression level (21). A recent study on the effects of different types of mutations in *Salmonella enterica* has shown that up to 25% of large deletions could result in a fitness increase, although the benefit of losing a

Significance

Evolution of microbes is dominated by horizontal gene transfer and the incessant host–parasite arms race that promotes the evolution of diverse antiparasite defense systems. The evolutionary factors governing these processes are complex and difficult to disentangle, but rapidly growing genome databases provide ample material for testing evolutionary models. Rigorous mathematical modeling of evolutionary processes, combined with computer simulation and comparative genomics, allowed us to elucidate the evolutionary regimes of different classes of microbial genes. Only genes involved in key informational and metabolic pathways are subject to strong selection, whereas most of the others are effectively neutral or even burdensome. Mobile genetic elements and defense systems are costly, supporting the understanding that their evolution is governed by the same factors.

Author contributions: J.I. and E.V.K. designed research; J.I., J.A.C., S.M., and M.I.K. performed research; J.I. and E.V.K. analyzed data; and J.I. and E.V.K. wrote the paper.

Reviewers: S.M., University of Illinois at Urbana–Champaign; and D.V., Columbia University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

¹To whom correspondence should be addressed. Email: koonin@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1704925114/-DCSupplemental.

particular gene critically depends on the environment (19). These findings emphasize the importance of averaging across multiple environmental conditions when it comes to estimating the fitness contribution of a gene. For the purpose of evolutionary analyses, a meaningful proxy for such an average can be obtained by inferring selection coefficients directly from the gene family abundances observed in large collections of genomes. The main difficulty in this case is disentangling the effects of selection from the effects of intrinsic loss bias, which normally requires a priori knowledge of the effective population size or the gene gain and loss rates (14, 22, 23).

Here we combine mathematical modeling, comparative genomics, and data compiled from mutation accumulation experiments to infer the characteristic contributions of selection and intrinsic DNA loss for different gene categories. To disentangle selection and loss bias, we first obtained an exact, time-dependent solution of the linear duplication–transfer–loss model with selection that governs the dynamics of gene copy numbers in a population of genomes (24–28). When applied to a large genomic data set, the model provides maximum likelihood estimates of the neutral equivalent (effective) loss bias, a composite parameter that amalgamates the effects of intrinsic loss bias (the loss bias before the action of selection) and selection. The selection coefficient can be extracted from the effective loss bias as long as the rate of gene loss is known, for which we used estimates from mutation accumulation experiments.

Our results show that with the exception of genes involved in core informational processes, most gene families are neutral or only slightly beneficial in the long term. Among the genetic elements that are typically considered parasitic, prophages show the highest fitness cost, followed by conjugative plasmids and transposons, which are only weakly deleterious in the long term. Notably, genes involved in antiparasite defense do not seem to provide long-term benefits on average but rather are slightly deleterious, almost to the same extent as transposons. We complete our analysis with an evaluation of the causes that make transposon dynamics qualitatively different from those of other gene classes and explore the effect of genome size on the rates of HGT, gene duplication, and gene loss.

Results

Duplication–Transfer–Loss Model of Gene Family Evolution. To describe the dynamics of a gene family size (gene copy number) in a population of genomes, we used a linear duplication–transfer–loss model with selection. Within a genome, the gene copy number can increase via duplication of the extant copies, which occurs at rate d per copy, or through the arrival of a new copy via HGT, at rate h independent on the copy number. Likewise, gene loss at rate l per copy leads to a decrease in the copy number. Duplication, HGT, and gene loss define a classical birth–death–transfer model at the genome level (24–27, 29). Selection is introduced through a contribution s to the fitness of a genome (s is positive for beneficial genes and negative for costly genes), which is multiplied by the gene copy number k . Specifically, we assume that fitness is additive, there is no epistasis, and the fitness contributions of all genes from the same family are the same. At the cell population level, the number of genomes carrying k copies, n_k , obeys the following system of differential equations:

$$\begin{aligned}\frac{dn_0}{dt} &= (g - h) n_0 + l n_1 \\ \frac{dn_k}{dt} &= (g - h - k(d + l - s)) n_k + (k + 1) l n_{k+1} + (h + d(k - 1)) n_{k-1}.\end{aligned}\quad [1]$$

The basal growth rate g was included for completeness, although it does not affect the copy number distribution. Moreover, the entire system can be restated in terms of the ratios of each of the parameters to the loss rate (see *SI Appendix* for more details). The linear

duplication–transfer–loss model with selection can be exactly solved for arbitrary initial conditions by formulating Eq. 1 as a first-order partial differential equation for the generating function and applying the method of characteristics (*SI Appendix*) (30, 31). The result is the copy number distribution, i.e., the fraction p_k of hosts with an arbitrary number of copies k at any time. In the case of a population where the gene family is initially absent, we obtain

$$p_k(t) = C(t) \frac{\left(\frac{d}{l} R(t)\right)^k}{k!} \frac{\Gamma(k + \frac{h}{d})}{\Gamma(\frac{h}{d})}, \quad [2]$$

with

$$R(t) = \frac{1 - e^{-(d/l) a^{-a^{-1}} t}}{(d/l) a - a^{-1} e^{-(d/l) a^{-a^{-1}} t}}$$

and

$$a = \frac{d - s + l + \sqrt{(d - s + l)^2 - 4dl}}{2d}. \quad [3]$$

In these expressions, time is measured in units of loss events, and $C(t)$ is a normalization factor that ensures that the sum of p_k over all k is equal to 1. A notable property of this solution is that as the system approaches the stationary state, selection, duplication, and loss merge into the composite parameter a , which, in the absence of selection, coincides with the inverse of the duplication/loss ratio (see *SI Appendix* for more details). Therefore, we refer to a^{-1} as the “neutral equivalent” (henceforth “effective”) duplication/loss ratio (d/l_e). It is also possible to define the effective HGT/loss ratio ($\frac{h}{l_e} = \frac{h}{d} \frac{d}{l_e}$) such that gene families with the same effective ratios have the same stationary distributions. The fitness contribution of a gene (i.e., selection to loss ratio) can be expressed in terms of the gene’s effective duplication/loss ratio and the actual (intrinsic) duplication/loss ratio as

$$\frac{s}{l} = \frac{\left(1 - \frac{d}{l_e}\right) \left(\frac{d}{l_e} - \frac{d}{l}\right)}{\frac{d}{l_e}}. \quad [4]$$

Duplication, Loss, and Selection in Different Functional Categories of Genes. We used the COUNT method (24) to estimate the effective duplication/loss ratio (d/l_e) associated to different gene families [defined as clusters of orthologous groups (COGs)] in 35 sets of closely related genomes [alignable tight genomic clusters (ATGCs)], which jointly encompass 678 bacterial and archaeal genomes (32, 33). As shown in the preceding section, the effective duplication/loss ratio (d/l_e) is a composite parameter that results from selection on gene copy number affecting the fixation of gene duplications and gene losses. For a neutral gene family, the effective duplication/loss ratio is simply the same as the ratio between the rates of gene duplication and gene loss. Because selection prevents the loss of beneficial genes, the effective duplication/loss ratios associated with beneficial genes are greater than their intrinsic duplication/loss ratios, whereas the opposite holds for genes (e.g., parasitic elements) that are costly to the host and tend to be eliminated by selection. Technically, the duplication term includes not only bona fide duplications but any process that causes an increase in copy number that is proportional to the preexisting copy number. Thus, HGT can also contribute to the duplication term in clonal populations, where the copy numbers of donors and recipients are highly correlated. Fig. 1A shows the effective duplication/loss ratios for gene families that belong to different functional categories [as defined under the COG classification (34)], as well as genes of transposons, conjugative plasmids, and prophages. For the majority of the gene families, the effective duplication/loss ratios are below 1, which is

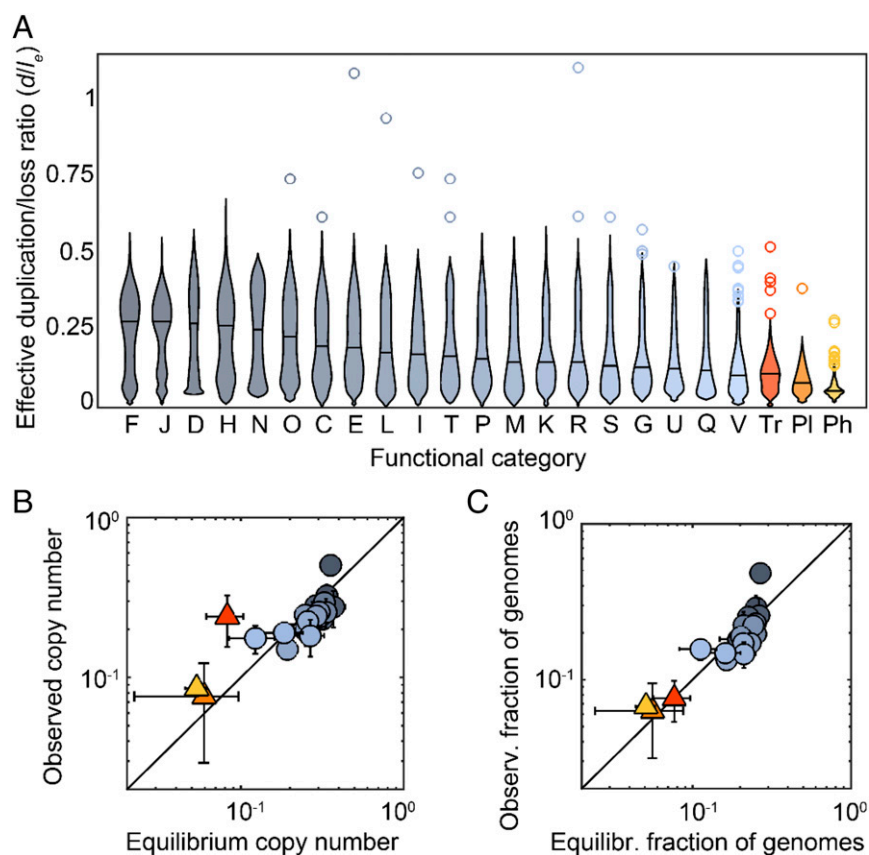


Fig. 1. Effective loss bias and mean abundances of gene families from different functional categories. (A) Distribution of the effective duplication/loss ratio d/l_e . Black horizontal lines indicate the median of each category. Outliers are represented as circles. Designations of the functional categories (modified from ref. 8): C, energy production and conversion; D, cell division; E, amino acid metabolism and transport; F, nucleotide metabolism and transport; G, carbohydrate metabolism and transport; H, coenzyme metabolism; I, lipid metabolism; J, translation; K, transcription; L, replication and repair; M, membrane and cell wall structure and biogenesis; N, secretion and motility; O, posttranslational modification, protein turnover, and chaperone functions; P, inorganic ion transport and metabolism; Q, biosynthesis, transport, and catabolism of secondary metabolites; R, general functional prediction only (typically, prediction of biochemical activity); S, function unknown; T, signal transduction; U, intracellular trafficking and secretion; V, defense mechanisms; Tr, transposon; PI, conjugative plasmid; and Ph, prophage or phage-related. Two extreme outliers, one from the transposons (transposase IS1595, $d/l_e = 1.4$) and one from category V (multidrug efflux pump subunit AcrB, $d/l_e = 1.6$), are not represented. (B) Comparison of the global (observed) mean copy number per family and the equilibrium copy number predicted by the model. Data points correspond to medians across functional categories (colors as in A; triangles are used to highlight genetic parasites). Error bars represent the 95% confidence interval for the median. The solid line corresponds to a perfect match between predictions and observations. The Spearman's correlation coefficients including and excluding parasites are $\rho = 0.80$ and 0.81 , respectively ($P < 10^{-4}$). (C) Fraction of genomes in which a family is present, compared with the expected fraction at equilibrium (Spearman's $\rho = 0.87$ and 0.80 , including and excluding parasites, $P < 10^{-4}$). Data points and error bars as in B.

compatible with the pervasive bias toward gene loss combined with (near) neutrality of numerous genes. In agreement with the notion that selection affects the effective duplication/loss ratios, their values decrease from the essential functional categories, such as translation and nucleotide metabolism, to the nonessential and parasitic gene classes. The apparent bimodality of the distributions for some functional categories (Fig. 1A) is likely due to their biological heterogeneity. For example, category N (secretion and motility) sharply splits into two major groups of gene families: (i) components of the flagellum and (ii) proteins involved in cellulose production and glycosyltransferases, with high d/l_e values for the former and much lower values for the latter (SI Appendix, Table S1).

The average fitness contribution of a gene can be inferred from its effective duplication/loss ratio provided that the intrinsic duplication/loss ratio is known (see preceding section). To estimate the intrinsic duplication/loss ratio (d/l), we used two independent approaches. The first approach was based on the assumption that a substantial fraction of genes from nonessential, but not parasitic, functional categories are effectively neutral. Considering that gene families in those categories are relatively well represented across taxa (we required them to be present in at least three different

ATGCs) and are not regarded as part of the mobilome (11), we would expect that, if not neutral, they are slightly beneficial and provide an upper bound for the intrinsic duplication/loss ratio. After sorting nonparasitic functional categories by their effective duplication/loss ratios (Fig. 1A), category K (transcription) was selected as the last category whose members arguably exert a positive average fitness effect. The intrinsic duplication/loss ratio was then calculated as the median of the effective duplication/loss ratios among the pool of gene families involved in poorly understood functions (R and S), carbohydrate metabolism (G), secretion (U), secondary metabolism (Q), and defense (V). In the second approach, we identified genes that are represented by one or more copies in a single genome, while absent in all other genomes of the same ATGC. Such genes [henceforth ORFans (35, 36)] are likely of recent acquisition and can be assumed neutral, if not slightly deleterious. The maximum likelihood estimate of the duplication/loss ratio obtained for ORFans provides, therefore, a lower bound for the intrinsic duplication/loss ratio (Methods and SI Appendix). The ratios obtained with both approaches were 0.124 [95% confidence interval (CI) 0.117 – 0.131] and 0.126 (95% CI 0.115 – 0.137). The two independent estimates are strikingly consistent with each other and

Table 1. Contributions of selection and the duplication/loss ratio to the evolution of different functional categories of genes and mobile elements

	d/l_e	s/l	$s (\times 10^{-8})$	
			Lower	Upper
F, nucleotide metabolism and transport	0.273	0.39	0.20	1.58
J, translation	0.273	0.39	0.20	1.58
D, cell division	0.266	0.39	0.19	1.56
H, coenzyme metabolism	0.260	0.38	0.19	1.54
N, secretion and motility	0.247	0.37	0.19	1.49
O, posttranslational modification, protein turnover, and chaperone functions	0.223	0.34	0.17	1.37
C, energy production and conversion	0.197	0.29	0.15	1.18
E, amino acid metabolism and transport	0.187	0.27	0.14	1.08
L, replication and repair	0.172	0.23	0.11	0.91
I, lipid metabolism	0.166	0.20	0.10	0.82
T, signal transduction	0.159	0.18	0.09	0.72
P, inorganic ion transport and metabolism	0.150	0.14	0.07	0.57
M, membrane and cell wall structure and biogenesis	0.140	0.09	0.05	0.36
K, transcription	0.140	0.09	0.05	0.36
R, general functional prediction only	0.140	0.09	0.04	0.36
S, function unknown	0.128	0.02	0.01	0.09
G, carbohydrate metabolism and transport	0.123	-0.02	-0.01	-0.07
U, intracellular trafficking and secretion	0.122	-0.02	-0.01	-0.09
Q, biosynthesis, transport, and catabolism of secondary metabolites	0.112	-0.10	-0.05	-0.40
V, defense	0.106	-0.16	-0.08	-0.62
V(i), antibiotic/drug resistance	0.135	0.06	0.03	0.25
V(ii), antipathogen defense	0.059	-1.05	-0.52	-4.18
Tr, transposon	0.104	-0.18	-0.09	-0.74
Pl, conjugative plasmid	0.079	-0.53	-0.27	-2.12
Ph, (pro)phage	0.047	-1.56	-0.78	-6.23

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.

robust to small changes in the methodology (SI Appendix). Accordingly, we took the average $d/l = 0.125$ as the intrinsic duplication/loss ratio. This value quantifies the intrinsic bias toward gene loss once the effect of selection is removed.

Quantitative estimates of the ratio between the selection coefficient and the loss rate (s/l) for each functional category are readily obtained by applying Eq. 4 to the effective duplication/loss ratios (Table 1). In the case of costly gene families, the ratio s/l quantifies the relative contributions of selection and loss in controlling the gene copy number. However, quantitative assessment of the selection coefficients from the s/l ratio requires knowledge of the intrinsic rates of gene loss in prokaryotic genomes. A compilation of published data from mutation accumulation experiments shows that disruption of gene coding regions due to small indels and/or large deletions occurs at rates between 5×10^{-9} and 4×10^{-8} per gene per generation (37–45), which yields the ranges for the selection coef-

ficients listed in Table 1. Assuming the effective size of typical microbial populations to fall between 10^8 and 10^9 (21, 46, 47), the selection coefficients yielded by these estimates indicate evolution determined by positive fitness contribution ($N_e s \gg 1$) for information processing categories (translation and replication) as well as some metabolic categories (especially nucleotide metabolism) and cellular functions (cell division and chaperones), an effectively neutral evolutionary regime for several categories including transcription, and evolution driven by negative fitness contribution ($N_e s \ll -1$) for defense genes and mobile genetic elements.

To shed light on the causes that make the defense genes slightly deleterious, we split the gene families in this category into two subcategories: (i) drug and/or antibiotic resistance and detoxification and (ii) restriction modification, CRISPR-Cas, and toxin-antitoxin. The median fitness effect substantially and significantly differs in sign and magnitude between both groups, with $s = (3.1 \times 10^{-10}, 2.5 \times 10^{-9})$ for genes involved in detoxification and drug resistance and $s = (-4.2 \times 10^{-8}, -5.2 \times 10^{-9})$ for genes involved in antiparasite defense (Mann-Whitney test, $p < 10^{-7}$). Thus, the drug resistance machinery is close to neutral whereas the antiparasite defense systems are about as deleterious as plasmids and somewhat more so than transposons. Among the latter, toxin-antitoxins are the most deleterious, followed by CRISPR-Cas and restriction modification, although the pairwise differences are only significant between toxin-antitoxins and restriction modification [$s = (-8.8 \times 10^{-8}, -1.1 \times 10^{-8})$ and $s = (-2.1 \times 10^{-9}, -2.6 \times 10^{-10})$, respectively; Mann-Whitney test, $p = 0.02$].

Long-Term Gene Dynamics and Bursts of Transposon Proliferation.

The loss biases and selection coefficients in Table 1 describe the dynamics of genes in groups of closely related genomes, with evolutionary distances of ~ 0.01 – 0.1 fixed substitutions per base pair. To investigate whether the same values apply at larger phylogenetic scales, we pooled data from all ATGCs and compared the global abundances of genes from different categories with the long-term equilibrium abundances expected from the model (Fig. 1B and C). In most categories, the observed copy number agrees with the predicted value, and the same holds for the fraction of genomes that harbor a given gene family.

Two notable exceptions are the genes involved in translation (category J) and the transposons. In the case of translation-related genes, the observed copy number is $\sim 40\%$ greater than expected (median observed 0.50, median expected 0.36, Wilcoxon test $p < 10^{-20}$), and the fraction of genomes with at least one copy is $\sim 80\%$ greater than expected (median observed 0.48, median expected 0.27, Wilcoxon test $p < 10^{-20}$). Such deviations reflect the inability of the model to reproduce a scenario in which selection acts to maintain a single member of most of the gene families in almost every genome, as is the case for translation. In the case of transposons, there is a dramatic excess of $\sim 213\%$ in the mean copy number (median observed 0.25, median expected 0.08, Wilcoxon test $p < 10^{-6}$) but no significant deviation in the fraction of genomes that carry transposons. Such excess of copies apparently results from occasional proliferation bursts that offset the prevailing loss-biased dynamics. Indeed, $\sim 12\%$ of the lineage-specific families of transposons show evidence of recent expansions, as indicated by effective duplication/loss ratios greater than 1, whereas the fraction of such families drops below 4% in other functional categories (Fig. 24, orange bars). Analysis of the typical burst sizes also reveals differences between transposons, with a mean burst size close to 4, and the rest of genes, with mean burst sizes around 2 (Fig. 24, gray line). Episodes of transposon proliferation are not evenly distributed among taxa but rather concentrate in a few groups, such as *Sulfolobus*, *Xanthomonas*, *Francisella*, and *Rickettsia* (Fig. 2B). The high prophage burst rate in *Xanthomonas* is due to the presence of a duplicated prophage related to P2-like viruses in *Xanthomonas citri*.

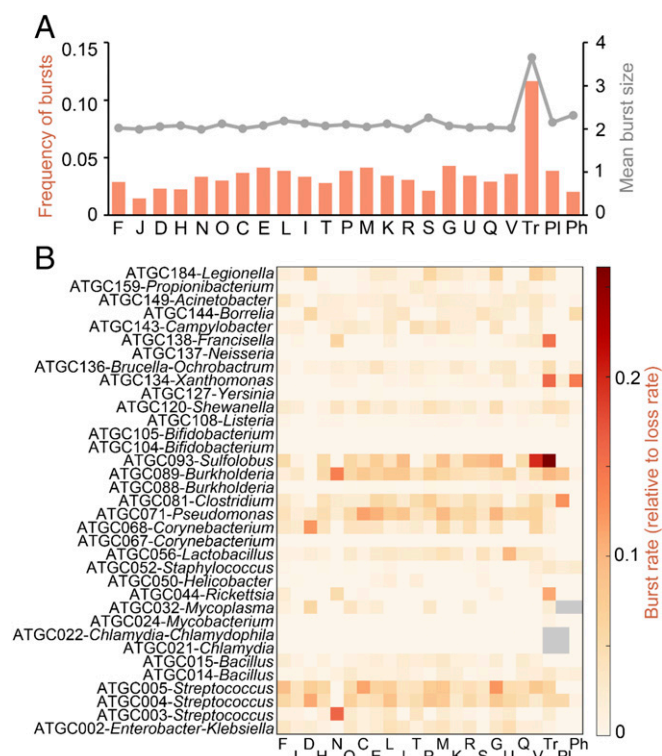


Fig. 2. Frequency and distribution of proliferation bursts in different functional categories of genes. (A) Orange (left axis) shows frequency of proliferation bursts, defined as the fraction of ATGC-COGs with effective duplication/loss ratio $d/l_e > 1$, split by functional category. Gray (right axis) shows mean burst size for these ATGC-COGs. (B) Burst rates in different ATGCs and functional categories, relative to the rate of gene loss. Designations of functional categories are the same as in Fig. 1 and Table 1.

To test whether the burst dynamics observed for transposons could explain the deviation in their global abundance, we analyzed a modified version of the model in which long phases of genome decay are punctuated by proliferative bursts of size K . Specifically, each decay phase was modeled as a duplication–transfer–loss process with selection, with initial condition $p_K = 1$, $p_{k \neq K} = 0$. Bursts occur at exponentially distributed intervals with the rate ϕ (note that $T = 1/\phi$ is the characteristic interval between two consecutive bursts). When a burst occurs, the duplication–transfer–loss process is reset to its initial condition. In this model, the time-extended average for the mean copy number, $\langle k \rangle$, becomes $\langle k \rangle = \int_0^\infty dt \sum_k k p_k(t) \phi e^{-\phi t}$. Using this expression it is possible to evaluate the expected mean copy number for any given value of the burst rate and the burst size (see *SI Appendix* for details). In the case of transposons, the fraction of families with signs of recent expansions leads to the estimate $\phi = 0.04$ (i.e., one burst for every 25 losses; *Methods*). For this burst rate, the modified model recovers the observed mean copy number if the burst size is set to $K = 4.2$, which is notably close to the value $K = 3.9$ estimated from the data.

Relationships Between Genome Size and Gene Duplication, Horizontal Transfer, and Loss Rates. We further investigated the relationships between the genome size and the factors that determine gene abundances. For each set of related genomes, we estimated the intrinsic duplication/loss ratio (d/l) and the total HGT/loss ratio (h/l) for genes from neutral categories and compared those to the mean genome size, quantified as the number of ORFs in the genome. As shown in Fig. 3, d/l is independent of the genome size, whereas h/l positively correlates with the genome size.

The same trends are confirmed by the analysis of ORFan abundances. Provided that the duplication rate is small compared with the loss rate, the number of ORFan families per genome constitutes a proxy for the ratio h/l . On the other hand, the fraction of ORFan families with more than one copy is a quantity that only depends on the ratio d/l (*SI Appendix*). As in the case of neutral gene families, the study of ORFans reveals a strong positive correlation between genome size and h/l but lack of significant correlation with d/l .

Because in prokaryotes genome size positively correlates with the effective population size (N_e) (14), we also explored the correlations between N_e and the ratios h/l and d/l (*SI Appendix*). The same qualitative correlations were detected; that is, h/l positively correlates with N_e , whereas d/l shows no correlation. However, the association between h/l and N_e becomes nonsignificant when genome size and N_e are jointly considered in an analysis of partial correlations. Therefore, it seems that the association between N_e and h/l is a by-product of the intrinsic correlation between effective population size and genome size.

Disentangling Environmental and Intrinsic Contributions to Fitness.

Because our estimates of the selection coefficients constitute ecological and temporal averages, a low selection coefficient might result not only from a genuine lack of adaptive value but, perhaps more likely, from the limited range of environmental conditions in which the given gene becomes useful. To disentangle the two scenarios, we compared the nonsynonymous to synonymous nucleotide substitution ratios (dN/dS) for different gene categories. The expectation is that genes that perform an important function in a rare environment would be characterized by low average selection coefficients (frequent loss) combined with intense purifying selection at the sequence level (low dN/dS) in those genomes that harbor the gene. Gene sequence analysis shows that in most cases, the dN/dS of a gene is primarily determined by the ATGC rather than by the functional category (*SI Appendix*, Fig. S1). These observations are compatible with the results of a previous analysis indicating that the median dN/dS value is a robust ATGC-specific feature (15). Notable exceptions are transposons and prophages, which show a high dN/dS in most taxa.

After accounting for the ATGC-related variability, we found a significant negative correlation between the selection coefficient of a functional class and the dN/dS (Fig. 4; Spearman's $\rho = -0.58$, $P = 0.004$). Such a connection between the selection pressures on gene dynamics and sequence evolution is to be expected under the straightforward assumption that genes that are more important for organism survival are subject to stronger selection on the sequence level and has been observed previously (48). However, genes involved in metabolic processes, especially carbohydrate metabolism, have lower dN/dS values than predicted from the overall trend (Fig. 4), suggesting that the effective neutrality of such genes results from the heterogeneity of environmental conditions. Among the gene categories with low selection coefficients, the dN/dS values of transposons, prophages, and gene families with poorly characterized functions are significantly greater than expected from the general trend, which is consistent with the notion that these genes provide little or no benefit to the cells that harbor them.

Gene Dynamics and Microbial Lifestyles. In an effort to clarify the biological underpinnings of the gene dynamics, we compared the effective duplication to loss ratios in microbes with three lifestyles: free-living, facultative host-associated, and obligate intracellular parasite (Fig. 5). In the first two groups, d/l_e drops from essential functional categories to nonessential categories and genetic parasites, with significantly higher values in free-living microbes than in facultative host-associated bacteria. Obligate intracellular parasites have remarkably low d/l_e values, as could be expected from their strong genomic degeneration. Notably, genetic parasites and genes from the defense category show the highest d/l_e among the genes of

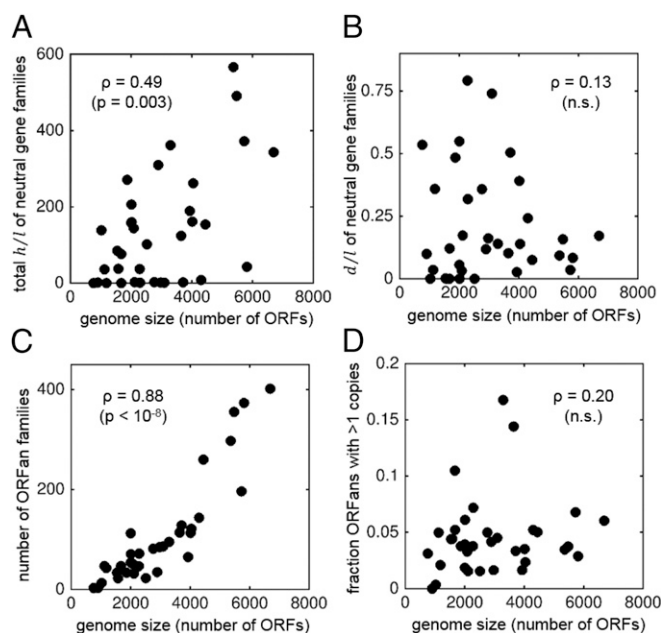


Fig. 3. Correlations between the genome size and potentially relevant parameters of gene family dynamics and genome architecture. Each point represents an ATGC. (A) Total HGT to loss ratio for genes from neutral categories. (B) Duplication to loss ratio for genes from neutral categories (both duplication and loss rates are calculated per copy). (C) Number of ORFan families per genome, which is an independent proxy for h/l . (D) Fraction of ORFan families with more than one copy, which is proportional to d/l . In each panel, the Spearman's ρ and significant P values are shown; nonsignificant (n.s.) P values are greater than 0.2.

intracellular parasites, although due to the small number of intracellular parasites in our dataset (only three ATGCs, with most genetic parasites restricted to the ATGC044 encompassing *Rickettsia*), this result must be taken with caution. We estimated the selection coefficients for free-living and facultative host-associated microbes, under the assumption that the intrinsic d/l is universally the same across the microbial diversity. The significant difference in d/l_e between the two lifestyles translates into consistently higher s values for most functional categories of genes in free-living microbes (SI Appendix, Fig. S2). Thus, the beneficial effects of most genes appear to be significantly greater in free-living compared with facultative host-associated bacteria, and in both these categories of microbes, selection for gene retention is dramatically stronger than it is in obligate, intracellular parasites.

Discussion

Multiple variants of the duplication–transfer–loss model and related multitype branching processes have been widely used to study the evolution of gene copy numbers (24, 25, 28, 49), especially in the context of transposons and other genetic parasites (22, 23, 26, 27, 50). To make the models tractable, most studies make simplifying assumptions, such as stationary state, absence of duplication, or lack of selection, and obtain the model parameters from the copy number distributions observed in large genomic datasets, relying on the assumption that model parameters are homogeneous across taxa. Here we derived an exact solution for the time-dependent duplication–transfer–loss model with additive selection and found that in general, it is impossible to distinguish neutral and costly elements solely based on the copy number distributions. This is the case because the effects of selection and loss bias blend into a composite parameter that is equivalent to an effective loss bias in a neutral scenario. Using the solution of the complete model, we investigated the copy number dynamics of a large number of gene

families in groups of related genomes, without the need to assume homogeneity of the HGT, duplication, and loss rates across taxa (8). We then used the expression that relates the parameter values under selection with their neutral equivalents to estimate the selection coefficients for different classes of genes.

The results of this analysis rely on several assumptions. First, the duplication–transfer–loss model was solved in a regime of linear selection that assumes that the benefit or cost of a gene family linearly grows with the gene copy number. This choice of the cost function, which is arguably suitable for genetic parasites, might be violated by ensembles of genes involved in processes that require tight dosage balance among the respective proteins, such as the translation system (51). For such genes, the fitness benefit will be underestimated because the observed number of family members is lower than predicted by the model. Second, to calculate the intrinsic loss bias (d/l), we assumed that certain classes of genes are effectively neutral. In that regard, two independent approaches were explored: (i) using ORFans as the neutral class and (ii) inferring the neutral categories based on plausible dispensability and a low position in the effective loss bias ranking. Notably, nearly identical values were obtained through both approaches, indicating that our estimates are robust to the choice of the neutral reference group. Third, the model assumes that duplication and deletion rates, as well as selection coefficients, are constant in time. It has been proposed that recently duplicated genes are subject to significantly higher loss rates and lower selection coefficients than older paralogs (52, 53). Should that be the case, recently duplicated gene copies would be short-lived, and their existence would not affect the generality of our results, provided that the duplication to loss ratio is understood as an effective parameter that accounts for the survival probability of a paralog beyond the initial phase. Finally, to convert the selection to loss ratios (s/l) to selection coefficients (s), we used two estimates of the loss rate l . A conservative estimate $l = 5 \times 10^{-9}$ was taken from the experimental study of medium to large deletions (in the range of 1 to 202 kb) in *Salmonella enterica* (37). Because small indels also contribute to the loss of genes via pseudogenization, we additionally considered a second, upper bound estimate, $l = 4 \times 10^{-8}$, which is the geometric mean of the indel rates collected from multiple mutation accumulation experiments (38–45) multiplied by an average target size of 1 kb per ORF.

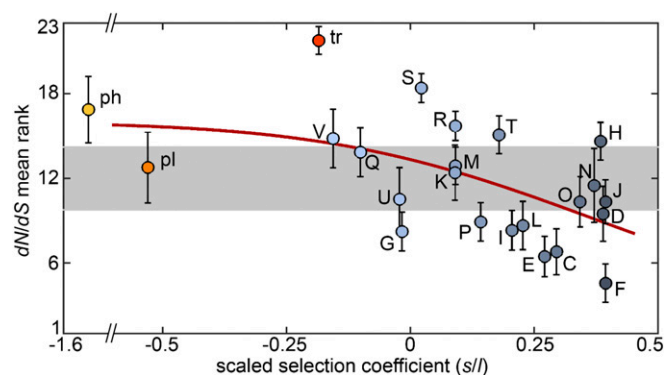


Fig. 4. Comparison between the scaled selection coefficients (s/l) of different functional categories and their characteristic nonsynonymous to synonymous mutation ratios (dN/dS). To account for ATGC-related variation, the dN/dS ratios for all categories within an ATGC were converted into ranks. Circles represent the mean ranks averaged across ATGCs, and error bars represent the SEM. Colors are the same as in Fig. 1. The horizontal gray band shows the theoretical 95% CI for the means of a null model where all categories have similar dN/dS (points above/below this interval indicate that the dN/dS of a category is significantly higher/lower than the expectation under the null model). The trend line (red) was obtained by fitting a monotonic spline curve to the data.

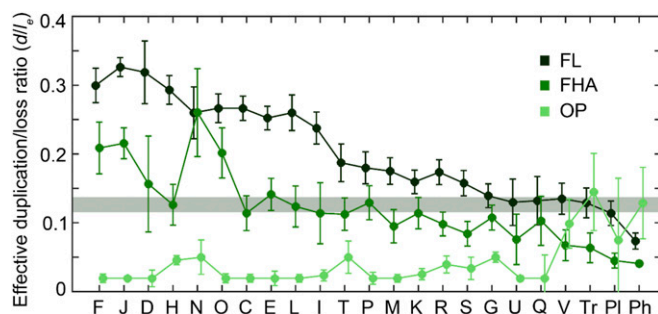


Fig. 5. Effective duplication to loss ratio (d/l_e) in free-living (FL), facultative host-associated (FHA), and obligate intracellular parasitic (OP) microbes. The designations of functional classes in the x axis are the same as in Fig. 1 and Table 1. The shaded band indicates the 95% CI for the intrinsic d/l_e estimated from neutral categories and ORFans. Error bars denote the 95% CI for the median d/l_e .

Our estimates yielded a broad range of selection coefficients that reflects positive, near zero (neutral) or negative fitness contributions of the respective genes. Notably, the ranking of the gene categories by fitness contribution is closely similar to the ranking by evolutionary mobility (gene gain and loss rates) (8) such that genes with positive fitness contributions are the least mobile. In accordance with the intuitive expectation, gene families involved in essential functions, in particular nucleotide metabolism and translation, occupy the highest ranks in the list of genes maintained by selection (highest positive s values; Table 1). The middle of the range of selection coefficients is occupied by functional categories of genes that are beneficial, sometimes strongly so, for microbes under specific conditions but otherwise could be burdensome, such as carbohydrate metabolism and ion transport. This inference was supported by analysis of selection on the protein sequence level that is reflected in the dN/dS ratio. Overall, we observed the expected significant negative correlation between the selection coefficient estimated from gene dynamics and dN/dS , indicating that functionally important genes are, on average, subject to strong constraints on the sequence level. However, for genes involved in metabolic processes, in particular carbohydrate metabolism, the dN/dS values are lower than expected given their average selection coefficients, which is consistent with relatively strong sequence-level selection in the subsets of microbes that have these genes. In agreement with this interpretation, when the s values for these categories were estimated separately for free-living and host-associated microbes, they turned out to be slightly beneficial in the former but costly in the latter.

In contrast, genetic parasites that negatively contribute to the fitness of the cell are at the bottom of the list of s values (Table 1). Among those, prophages are the most costly class, whereas plasmids and especially transposons evolve under regimes closer to neutrality. Prophages, plasmids, and transposons differ substantially in the magnitude of the associated selection coefficients: selection is strong and effective against prophages ($N_e s \sim -10$) and moderate against transposons and plasmids ($N_e s \sim -1$). These differences are consistent with the differences in the lifestyles between these selfish elements whereby transposons and plasmids are relatively harmless to the host cell, apart from being an energetic burden, whereas prophages have the potential to kill the host upon lysis (20, 54). Accordingly, genetic parasites also differ in the relative importance that selection and deletions play in keeping them under control. Both selection and deletions contribute to the removal of prophages (the contribution of selection being ~ 1.6 times greater), whereas deletion is the main cause of plasmid and transposon loss (roughly twice as important as selection for plasmids and 5 times as important in the case of transposons). The demonstration that transposons are only weakly selected

against and are lost primarily due to the intrinsic deletion bias is compatible with the wealth of degenerated insertion sequences found in many bacterial genomes (55–57). Conversely, deleterious elements, such as prophages, whose spread is limited by selection against high copy numbers, present fewer degenerated copies than lower cost elements, such as transposons.

One of the most interesting and, at least at first glance, unexpected observations made in the course of this work is that genes encoding components of antipathogen defense systems are on average deleterious, with an average cost similar to or even greater than the cost of plasmids (Table 1). In part, this is likely to be the case because some of the most abundant defense systems, such as toxin–antitoxins and restriction modification modules, clearly display properties of selfish genetic elements and, moreover, are additive to host cells (58–61). Indeed, in agreement with the partially selfish character of such defense modules, we found that toxin–antitoxins are the most deleterious category of genetic elements in microbes, apart from prophages. More generally, the patchy distribution of defense systems in prokaryotic genomes, together with theoretical and experimental evidence, suggests that defense systems incur nonnegligible fitness costs that are thought to stem primarily from autoimmunity and abrogation of HGT and, therefore, are rapidly eliminated when not needed (62–64).

Long-term transposon dynamics is well described by a model that combines long phases of decay, during which transposons behave as inactive genetic material, punctuated by small proliferation bursts that produce on average four new copies. Despite the simplicity of this model, it captures, at least qualitatively, the heterogeneity of transposition rates among transposon families (65) and environmental conditions (66, 67). Unlike large expansions, which are rare events typically associated with ecological transitions affecting the entire genome (68–71), small bursts occur frequently and affect a sizable fraction of transposon families. Some well-known instances of large transposon expansions become apparent in our analysis that identified taxa with unusually high burst rates, such as *Xanthomonas*, *Burkholderia*, and *Francisella*, in accord with previous observations (70, 71). In most other taxa, transposon decay is the dominant process, which is the expected trend, given that transposition is tightly regulated and a large fraction of transposon copies are inactive (72, 73). The small fitness cost of transposons in the decay phase is also consistent with a nonproliferative scenario, where the fitness effect is reduced to the energetic cost of replication and expression (21). Due to the rapidity of bursts, our methodology cannot be used to assess the cost of a transposon during the burst phase. Because active transposons likely impose a larger burden on the host (74), variation in burst sizes is likely to reflect differences in the intensity of selection and the duration of proliferative episodes.

Apart from the transposons, the only notable case of burst-driven dynamics corresponds to genes from the defense category in *Sulfolobus*. A closer inspection of this group reveals multiple instances of duplications, gains, and losses of CRISPR–Cas systems as also observed previously (75). In the case of prophages, the low burst rate is likely to reflect genuine lack of bursts or our inability to detect them due to the dominant, selection-driven fast decay dynamics. Indeed, given the fitness cost that we estimated for prophages, a burst of prophages would decay almost three times faster than a burst of transposons of similar size.

The effective size of microbial populations positively correlates with the genome size, which led to the hypothesis that the genome dynamics is dominated by selection acting to maintain slightly beneficial genes (14, 16). In the present analysis, when gene families from all functional categories are pooled, the median fitness contribution per gene is $N_e s \sim 0.1$, which provides independent support for this weak selection-driven concept of microbial genome evolution. In that framework, the fact that genetic parasites are more abundant in large genomes, as reported previously (76–78) and confirmed by our data, seemingly raises a

paradox: the same genomes where selection works more efficiently to maintain beneficial genes also harbor more parasites. A possible solution comes from our observation that the HGT to loss ratio (where the HGT rate is measured per genome and the loss rate is measured per gene) grows with the genome size. Such behavior, which had been already noted for transposons (23) and agrees with the recently derived genome-average scaling law (14), is likely to result, at least in part, from larger genomes providing more nonessential regions where a parasite can integrate without incurring major costs to the cell. Alternatively or additionally, the observed dependence could emerge if duplication and loss rates per gene decreased with genome size, whereas the HGT rate remains constant. Indeed, an inverse correlation between the genome size and the duplication and loss rates could be expected as long as mutation rates appear to have evolved to lower values in populations with larger N_e (41, 79).

Taken together, the results of this analysis reveal the relative contributions of selection and intrinsic deletion bias to the evolution of different classes of microbial genes and selfish genetic elements. Among other findings, we showed that the genome-averaged selection coefficients are low, and evolution is driven by strong selection only for a small set of essential genes. In addition, we detected substantial, systematic differences between the evolutionary regimes of bacteria with different lifestyles, with much stronger selection for gene retention in free-living microbes compared with parasites, especially obligate, intracellular ones. This difference appears to be fully biologically plausible in that diversification of the metabolic, transport, and signaling capabilities is beneficial for free-living microbes but not for parasites that therefore follow the evolutionary route of genome degradation.

Counterintuitive as this might be, we show that antiparasite defense systems are generally deleterious for microbes, roughly to the same extent as mobile elements. These results are compatible with the previously observed highly dynamic evolution of such systems that are kept by microbes either when they are essential to counteract aggressive parasites or due to their own selfish and addictive properties. These findings can be expected to foster further exploration of the interplay between genome size; effective population size; the rates of horizontal transfer; duplication and loss of genes; and the dynamics of mobile elements in the evolution of prokaryotic populations and, eventually, the entire microbial biosphere.

Methods

Gene Copy Number Dynamics. Let $n_k(t)$ be the number of genomes that carry k copies of the gene of interest at time t . We define the generating function $G(z, t) = \sum_k z^k n_k(t)$. In terms of the generating function, Eq. 1 becomes $\frac{\partial G}{\partial t} = (\rho z^2 - \alpha z + 1) \frac{\partial G}{\partial z} + \beta(z-1)G$, where $\rho = \frac{d}{\mu}$, $\beta = \frac{h}{\mu}$, and $\alpha = 1 + \rho - \xi$. This equation can be solved for any initial condition by applying the method of characteristics (SI Appendix). The generating function for the copy number distribution $p_k(t)$ is then obtained as $H(z, t) = G(z, t)/G(1, t)$. The explicit values of $p_k(t)$ are recovered as the coefficients of the series expansion of $H(z, t)$ with respect to z .

Estimation of the Effective Ratios d/l_e and h/l_e from Genomic Data. Genomic data were obtained from an updated version of the ATGC database that clusters genomes from bacteria and archaea into closely related groups (33). We analyzed 35 of the largest ATGCs (34 bacterial and 1 archaeal group) that included 10 or more genomes each. For each of those ATGCs, clusters of orthologous genes shared among genomes of the same ATGC (ATGC-COGs) were identified (33, 80), and rooted species trees were generated as described previously (8).

The effective duplication/loss ratio (d/l_e) and transfer/loss ratio (h/l_e) for each ATGC-COG were estimated with the software COUNT (24), which optimizes the

parameters of a duplication–transfer–loss model analogous to the model described above under the assumption of neutrality (81). The output of the program was postprocessed to obtain ATGC-COG-specific rates as described in ref. 20. ATGC-COGs were assigned to families based on their COG and pfam annotations. COG and pfam annotations were also used to classify families into functional categories. At the family level, the representative ratios d/l_e and h/l_e of a family were obtained as the median d/l_e and the sum of h/l_e , respectively, among its constituent ATGC-COGs. The mean copy number of a family was calculated as the average, across all ATGCs, of the ATGC-specific mean abundances (ATGC-COGs belonging to the same family in the same ATGC were pooled to obtain the ATGC-specific mean abundance, whereas the ATGC-specific mean for absent families was set to zero). The fraction of genomes that contain a family was calculated in a similar manner. This approach minimizes the bias associated to nonuniform ATGC sizes. To minimize inference artifacts associated to small families, only those families encompassing at least five ATGC-COGs from at least three ATGCs were considered for further analyses.

Estimation of the Intrinsic Duplication/Loss Ratio. Two approaches were used to estimate the intrinsic duplication/loss ratio d/l . In the first approach, putative neutral families from categories R, S, G, U, Q, and V were pooled, and the median d/l_e was chosen to serve as the estimate of d/l . The 95% confidence interval was calculated with the formula $\text{median} \pm 1.7 \times (1.25 \text{ IQR}/1.35 \sqrt{N})$, where IQR is the interquartile range and N is the number of families (82). In the second approach, the copy numbers of ATGC-COGs that are specific to one single genome were used to infer the ratio d/l under the assumption that such genes are of recent acquisition and effectively neutral. To that end we used the solution of the duplication–transfer–loss model to derive a maximum likelihood estimate of d/l given a list of single-genome ATGC-COGs, their copy numbers, and the time since the last branching event in the genome tree (in units of loss events, as provided by COUNT). Explicit formulas and their derivation are discussed in SI Appendix. Likelihood maximization was carried out using the Nelder–Mead simplex method as implemented in MATLAB R2016b. The 95% confidence interval was determined by the values of d/l whose log-likelihood was 1.92 units smaller than the maximum log-likelihood (83).

Burst Frequency, Rate, and Size. The frequency of bursts was calculated as the fraction of ATGC-COGs in which $d/l_e > 1$. The burst rate ϕ was estimated by maximum likelihood, assuming that bursts occur randomly at exponentially distributed intervals, such that the probability of observing a burst in a tree of phylogenetic depth t is equal to $1 - e^{-\phi t}$. Accordingly, the log-likelihood of observing n_{atgc} bursts in an ATGC with N_{atgc} ATGC-COGs is $\text{LogLk}_{\text{atgc}} = n_{\text{atgc}} \log(1 - e^{-\phi t_{\text{atgc}}}) - (N_{\text{atgc}} - n_{\text{atgc}}) \phi t_{\text{atgc}}$, where t_{atgc} is the depth of the ATGC tree in units of loss events (SI Appendix). The global log-likelihood is the sum of the contributions from all ATGCs. As a proxy for the burst size we used the maximum copy number observed in each ATGC-COG. For each category, the characteristic burst size was calculated as the quotient between the mean burst size in ATGC-COGs with $d/l_e > 1$ and the baseline defined by the mean of the maxima in the rest of ATGC-COGs.

Estimation of the Characteristic dN/dS Ratios. The dN/dS of every ATGC-COG was calculated as follows. Starting from the multiple sequence alignment, the program codeml from the PAML (phylogenetic analysis by maximum likelihood) package (84) was used to obtain the dN/dS for each pair of sequences in the ATGC-COG. The conditions $0.01 < dN < 3$ and $0.01 < dS < 3$ were used to select informative gene pairs. The representative dN/dS for the ATGC-COG was obtained as the median dN/dS among the informative pairs. In the next step, ATGC-COGs from the same ATGC that belong to the same functional category were pooled, and the median of their dN/dS was taken as the representative dN/dS . To account for ATGC-related effects, the dN/dS values of all categories within an ATGC were converted into ranks. The null hypothesis that all categories are equal in terms of their dN/dS was rejected by a Skillings–Mack test ($t = 939.7$, $df = 22$, $P < 10^{-20}$). To identify which categories significantly deviate from the null hypothesis, the mean rank of each category was compared with the theoretical 95% CI for the mean of 35 samples taken from a discrete uniform distribution in the interval from 1 to 23.

- Koonin EV (2011) *The Logic of Chance: The Nature and Origin of Biological Evolution* (FT Press, Upper Saddle River, NJ).
- Lynch M (2007) *The Origins of Genome Architecture* (Sinauer Associates, Sunderland, MA).
- Koonin EV, Wolf YI (2012) Evolution of microbes and viruses: A paradigm shift in evolutionary biology? *Front Cell Infect Microbiol* 2:119.
- Moran NA, Bennett GM (2014) The tiniest tiny genomes. *Annu Rev Microbiol* 68:195–215.

- Han K, et al. (2013) Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. *Sci Rep* 3:2101.
- Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol* 1:127–136.
- Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: The bacterial pan-genome. *Curr Opin Microbiol* 11:472–477.

8. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV (2014) Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Biol* 12:66.
9. Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* 36:6688–6719.
10. Wolf YI, Makarova KS, Lobkovsky AE, Koonin EV (2016) Two fundamentally different classes of microbial genomes. *Nat Microbiol* 2:16208.
11. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 3:722–732.
12. Lobkovsky AE, Wolf YI, Koonin EV (2013) Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol Evol* 5:233–242.
13. Kuo CH, Ochman H (2009) Deletional bias across the three domains of life. *Genome Biol Evol* 1:145–152.
14. Sela I, Wolf YI, Koonin EV (2016) Theory of prokaryotic genome evolution. *Proc Natl Acad Sci USA* 113:11399–11407.
15. Novichkov PS, Wolf YI, Dubchak I, Koonin EV (2009) Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. *J Bacteriol* 191:65–73.
16. Kuo CH, Moran NA, Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Res* 19:1450–1454.
17. Kuo CH, Ochman H (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genet* 6:e1001050.
18. Lee M-C, Marx CJ (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet* 8:e1002651.
19. Koskineniemi P, Sun S, Berg OG, Andersson DI (2012) Selection-driven gene loss in bacteria. *PLoS Genet* 8:e1002787.
20. Iranzo J, Puigbò P, Lobkovsky AE, Wolf YI, Koonin EV (2016) Inevitability of Genetic Parasites. *Genome Biol Evol* 8:2856–2869.
21. Lynch M, Marinov GK (2015) The bioenergetic costs of a gene. *Proc Natl Acad Sci USA* 112:15690–15695.
22. Bichsel M, Barbour AD, Wagner A (2013) Estimating the fitness effect of an insertion sequence. *J Math Biol* 66:95–114.
23. Iranzo J, Gómez MJ, López de Saro FJ, Manrubia S (2014) Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLOS Comput Biol* 10:e1003680.
24. Csürös M (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26:1910–1912.
25. Karev GP, Wolf YI, Berezovskaya FS, Koonin EV (2004) Gene family evolution: An in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol Biol* 4:32.
26. van Passel MWJ, Nijveen H, Wahl LM (2014) Birth, death, and diversification of mobile promoters in prokaryotes. *Genetics* 197:291–299.
27. Basten CJ, Moody ME (1991) A branching-process model for the evolution of transposable elements incorporating selection. *J Math Biol* 29:743–761.
28. Huynen MA, van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* 15:583–589.
29. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol Biol* 2:18.
30. Gardiner CW (2004) *Handbook of Stochastic Methods* (Springer-Verlag, Berlin).
31. van Kampen NG (2001) *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam).
32. Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I (2009) ATGC: A database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res* 37:D448–D454.
33. Kristensen DM, Wolf YI, Koonin EV (2017) ATGC database and ATGC-COGs: An updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation. *Nucleic Acids Res* 45:D210–D218.
34. Galperin MY, Makarova KS, Wolf YI, Koonin EV (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 43:D261–D269.
35. Siew N, Fischer D (2003) Unravelling the ORFan puzzle. *Comp Funct Genomics* 4: 432–441.
36. Yu G, Stoltzfus A (2012) Population diversity of ORFan genes in *Escherichia coli*. *Genome Biol Evol* 4:1176–1187.
37. Nilsson AI, et al. (2005) Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci USA* 102:12112–12116.
38. Dillon MM, Sung W, Lynch M, Cooper VS (2015) The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. *Genetics* 200:935–946.
39. Dillon MM, Sung W, Sebra R, Lynch M, Cooper VS (2017) Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Mol Biol Evol* 34:93–109.
40. Long H, et al. (2015) Background mutational features of the radiation-resistant bacterium *Deinococcus radiodurans*. *Mol Biol Evol* 32:2383–2392.
41. Sung W, et al. (2016) Evolution of the insertion-deletion mutation rate across the tree of life. *G3 (Bethesda)* 6:2583–2591.
42. Sung W, et al. (2015) Asymmetric context-dependent mutation patterns revealed through mutation-accumulation experiments. *Mol Biol Evol* 32:1672–1683.
43. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proc Natl Acad Sci USA* 109:18488–18492.
44. Sung W, et al. (2012) Extraordinary genome stability in the ciliate *Paramecium tetraurelia*. *Proc Natl Acad Sci USA* 109:19339–19344.
45. Dettman JR, Sztepanacz JL, Kassen R (2016) The properties of spontaneous mutations in the opportunistic pathogen *Pseudomonas aeruginosa*. *BMC Genomics* 17:27.
46. Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60:327–349.
47. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404.
48. Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* 13:2229–2235.
49. Novozhilov AS, Karev GP, Koonin EV (2006) Biological applications of the theory of birth-and-death processes. *Brief Bioinform* 7:70–85.
50. Moody ME (1988) A branching process model for the evolution of transposable elements. *J Math Biol* 26:347–357.
51. Veitia RA, Potier MC (2015) Gene dosage imbalances: Action, reaction, and models. *Trends Biochem Sci* 40:309–317.
52. Axelsen JB, Yan KK, Maslov S (2007) Parameters of proteome evolution from histograms of amino-acid sequence identities of paralogous proteins. *Biol Direct* 2:32.
53. Innan H, Kondrashov F (2010) The evolution of gene duplications: Classifying and distinguishing between models. *Nat Rev Genet* 11:97–108.
54. Jalasvuori M, Koonin EV (2015) Classification of prokaryotic genetic replicators: Between selfishness and altruism. *Ann N Y Acad Sci* 1341:96–105.
55. Cerveau N, Leclercq S, Leroy E, Bouchon D, Cordaux R (2011) Short- and long-term evolutionary dynamics of bacterial insertion sequences: Insights from *Wolbachia endosymbionts*. *Genome Biol Evol* 3:1175–1186.
56. Nelson WC, Wollerman L, Bhaya D, Heidelberg JF (2011) Analysis of insertion sequences in thermophilic cyanobacteria: Exploring the mechanisms of establishing, maintaining, and withstanding high insertion sequence abundance. *Appl Environ Microbiol* 77:5458–5466.
57. Brügger K, et al. (2002) Mobile elements in archaeal genomes. *FEMS Microbiol Lett* 206:131–141.
58. Van Melderden L (2010) Toxin-antitoxin systems: Why so many, what for? *Curr Opin Microbiol* 13:781–785.
59. Van Melderden L, Saavedra De Bast M (2009) Bacterial toxin-antitoxin systems: More than selfish entities? *PLoS Genet* 5:e1000437.
60. Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res* 29:3742–3756.
61. Furuta Y, Kobayashi I (2011) Restriction-modification systems as mobile epigenetic elements. *Bacterial Integrative Mobile Genetic Elements*, eds Roberts AP, Mullany P (Landes Bioscience, Austin, TX).
62. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, Koonin EV (2012) Viral diversity threshold for adaptive immunity in prokaryotes. *MBio* 3:e00456-12.
63. Koonin EV, Zhang F (2017) Coupling immunity and programmed cell suicide in prokaryotes: Life-or-death choices. *BioEssays* 39:1–9.
64. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV (2015) Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. *BMC Evol Biol* 15:43.
65. Sousa A, Bourgard C, Wahl LM, Gordo I (2013) Rates of transposition in *Escherichia coli*. *Biol Lett* 9:20130838.
66. Ohtsubo Y, Genka H, Komatsu H, Nagata Y, Tsuda M (2005) High-temperature-induced transposition of insertion elements in *Burkholderia multivorans* ATCC 17616. *Appl Environ Microbiol* 71:1822–1828.
67. Naas T, Blot M, Fitch WM, Arber W (1994) Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics* 136:721–730.
68. Beare PA, et al. (2009) Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins within the genus *Coxiella*. *Infect Immun* 77:642–656.
69. Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* 14:627–633.
70. Mira A, Pushker R, Rodríguez-Valera F (2006) The Neolithic revolution of bacterial genomes. *Trends Microbiol* 14:200–206.
71. Rohmer L, et al. (2007) Comparison of *Francisella tularensis* genomes reveals evolutionary events associated with the emergence of human pathogenic strains. *Genome Biol* 8:R102.
72. Nagy Z, Chandler M (2004) Regulation of transposition in bacteria. *Res Microbiol* 155: 387–398.
73. Filée J, Siguier P, Chandler M (2007) Insertion sequence diversity in archaea. *Microbiol Mol Biol Rev* 71:121–157.
74. Elena SF, Eklunwe L, Hajela N, Oden SA, Lenski RE (1998) Distribution of fitness effects caused by random insertion mutations in *Escherichia coli*. *Genetica* 102-103:349–358.
75. Garrett RA, et al. (2011) CRISPR-based immune systems of the *Sulfolobales*: Complexity and diversity. *Biochem Soc Trans* 39:51–57.
76. Zhou F, Olman V, Xu Y (2008) Insertion sequences show diverse recent activities in Cyanobacteria and Archaea. *BMC Genomics* 9:36.
77. Touchon M, Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* 24:969–981.
78. Touchon M, Bernheim A, Rocha EP (2016) Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J* 10:2744–2754.
79. Lynch M, et al. (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* 17:704–714.
80. Kristensen DM, et al. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* 26:1481–1487.
81. Csürös M, Miklós I (2006) A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. *Research in Computational Molecular Biology. RECOMB 2006. Lecture Notes in Computer Science*, eds Apostolico A, Guerra C, Istrail S, Pevzner PA, Waterman M (Springer, Berlin), Vol 3909.
82. McGill R, Tukey JW, Larsen WA (1978) Variations of box plots. *Am Stat* 32:12–16.
83. Hudson DJ (1971) Interval estimation from likelihood function. *J R Stat Soc B* 33:256–262.
84. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

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	2
S2 Analytic solution of the model	3
S2.1 General solution	3
S2.2 Solution when the gene is initially absent	4
S2.3 Solution when there are initially K copies per cell	4
S3 Distribution of the gene copy number	5
S3.1 Generating function	5
S3.2 Copy number distribution: explicit expression	5
S3.3 Average copy number	6
S3.3.1 Neutral genes ($\sigma = 0$)	6
S3.3.2 Deleterious genes ($\sigma < 0$)	6
S3.3.3 Beneficial genes ($\sigma > 0$)	7
S3.4 Fraction of genomes without the gene	7
S3.5 Characteristic time to equilibrium	7
S3.6 Multiple gene families with equivalent biological properties	7
S3.7 Copy number distribution when the initial population has K copies per cell	8
S4 Equivalence between neutrality and selection	9
S4.1 Equivalence of the asymptotic distributions	9
S4.2 Equivalence of transient distributions	10
S4.3 Recovery time after perturbations	10
S5 Maximum likelihood estimation of the duplication to loss ratio	11
S5.1 Likelihood function	11
S5.2 Practical implementation	12
S6 Model with proliferation bursts	13
S7 Correlations with effective population size	14
S8 Supplementary Figures	15
S9 Supplementary Tables	17

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:

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

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

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

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

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

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

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

The equation of the characteristic curves of (S5) is

$$\frac{dz}{dt} = -\rho z^2 + \alpha z - 1 \quad (\text{S6})$$

or equivalently

$$\frac{dz}{-\rho z^2 + \alpha z - 1} = dt \quad (\text{S7})$$

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

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

where

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

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

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

and solving for z ,

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

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

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

Equation (S12) admits a solution of the form

$$G(t) = F(\xi) W(t) \quad (\text{S13})$$

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

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

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

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

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

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

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

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

where

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

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

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

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

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

In the asymptotic limit $t \rightarrow \infty$ the function $R(t) \rightarrow (\rho a)^{-1}$ 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

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

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

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

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)

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

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.

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

where

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

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

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

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

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

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:

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

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 $\rho \rightarrow 0$, the asymptotic expression of the generating function becomes:

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

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

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

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

S3.3 Average copy number

The average copy number is obtained from the generating function as

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

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

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

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

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

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

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

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 $\rho \rightarrow 1$ we get $\langle k \rangle = \beta t$, and the gene copies accumulate in the genome at a constant rate.
- If $\rho < 1$ (loss bias), then $\langle k \rangle \rightarrow \frac{\beta}{1 - \rho}$, 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

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

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

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

In the asymptotic limit $t \rightarrow \infty$

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

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 equilibrium 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 $p_{k',b}$. 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 generating functions:

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

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 $i = 1, \dots, M$) one for each gene family. The generating function for the total copy number, H_T , is obtained from eqn. (S23) as

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

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

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

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:

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

where we defined

$$\begin{aligned} \psi &= \beta/\rho \\ \phi &= 1/a \end{aligned} \quad (\text{S42})$$

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

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

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

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

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:

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

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 $t \rightarrow \infty$. 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

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

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

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

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 expression 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:

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

The first term in the integral is equal to

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

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 :

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

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

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

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

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

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 $\sigma \rightarrow 0$, the result

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

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

where

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

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 parameter ρ becomes

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

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

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

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

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

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:

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

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

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

where

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

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

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

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, $\phi \ll \rho a - a'$, 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:

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

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

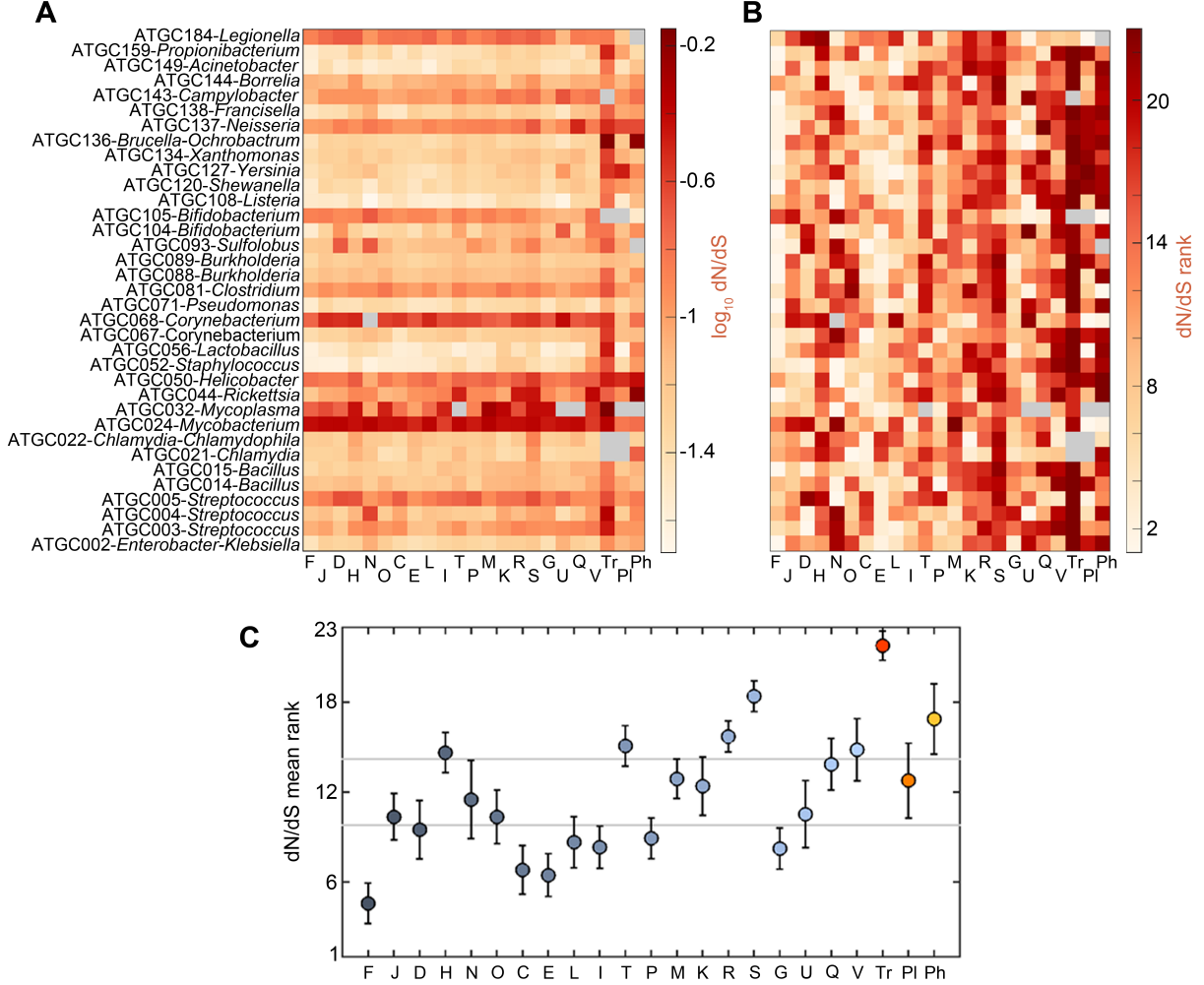


Figure S1. Non-synonymous to synonymous mutation ratio (dN/dS) for different gene categories. A: The color intensity represents, in logarithmic scale, the median dN/dS of the ATGC-COGs that belong to a specific ATGC and category. The cases where the number of informative pairs was too small to infer 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).

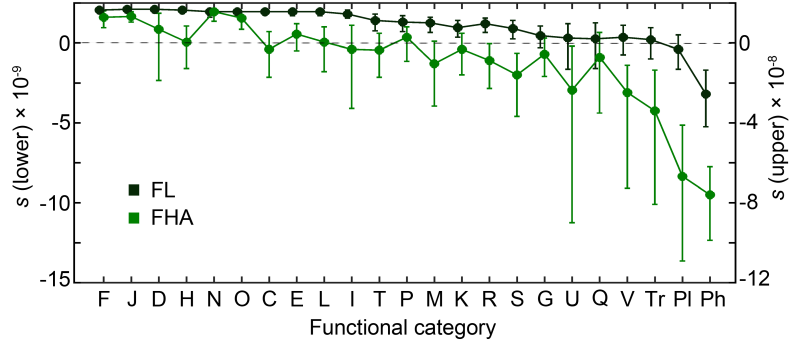


Figure S2. Selection coefficients in free-living (FL) and facultative host-associated (FHA) microbes, under the assumption that the intrinsic duplication to loss ratio (d/l) is the same in both lifestyles. The designations of functional classes in the x-axis are the same as in Figure 1 and Table 1. Error 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).

	d/l_e	s/l	$s (\times 10^{-8})$	
			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).