Disentangling the effects of selection and loss bias on gene dynamics

Jaime Iranzo[1], José A. Cuesta[2], Susanna Manrubia[3], Mikhail I. Katsnelson[4], Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; [2]Grupo Interdisciplinar de Sistemas Complejos (GISC); Departamento de Matemáticas, Universidad Carlos III de Madrid, Spain; Institute for Biocomputation and Physics of Complex Systems, Zaragoza, Spain; UC3M-BS Institute of Financial Big Data (IFiBiD); [3]Grupo Interdisciplinar de Sistemas Complejos (GISC); National Biotechnology Centre (CSIC), Madrid, Spain; [4]Institute for Molecules and Materials, Radboud University, Nijmegen, 6525AJ, Netherlands. *For correspondence: koonin@ncbi.nlm.nih.gov

ABSTRACT

We combine mathematical modelling 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 non-parasitic genes is low albeit positive, compatible with the 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 anti-parasite 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 the duplication to loss ratio, correlates with genome size, potentially explaining the increased abundance of neutral and costly elements in larger genomes.

SIGNIFICANCE

Evolution of microbes is dominated by horizontal gene transfer and the incessant host-parasite arms race that promotes the evolution of diverse anti-parasite defense systems. The evolutionary factors governing these processes are complex and difficult to disentangle but the 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.

/body

Introduction

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 110kb and 140 protein coding genes in the smallest intracellular symbionts (4) to almost 15Mb and more than 10,000 genes in the largest myxobacteria (5). Beyond a core of approximately 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 ORFans, 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 comprise 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 towards DNA deletion (and hence gene loss) that characterizes mutational processes in prokaryotes (as well as eukaryotes) results in non-adaptive genome 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, non-uniform 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 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 modelling, 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 prior to 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 anti-parasite 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 that 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 employed 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:

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

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 text 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 text)(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

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

with $(t) = \dfrac{1-e^{-((d/l)\,a-a^{-1})\,t}}{(d/l)\,a-a^{-1}\,e^{-((d/l)\,a-a^{-1})\,t}}$, and $a = \dfrac{d-s+l+\sqrt{(d-s+l)^2-4d}}{2d}$. 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 text 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 $\left(\frac{h}{l}_e = \frac{h}{d}\frac{d}{l}_e\right)$ 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

$$\frac{s}{l} = \frac{\left(1-\frac{d}{l}_e\right)\left(\frac{d}{l}_e - \frac{d}{l}\right)}{\frac{d}{l}_e} \qquad (3)$$

Duplication, loss and selection in different functional categories of genes

We used the COUNT method to estimate the effective duplication/loss ratio (d/l_e) associated to different gene families (defined as Clusters of Orthologous Groups, or COGs) in 35 sets of closely related genomes (Alignable Tight Genomic Clusters, or 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 compatible with the pervasive bias towards 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 non-essential 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 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 employed two independent approaches. The first approach was based on the assumption that a substantial fraction of genes from non-essential, 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 3 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 non-parasitic 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, 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 (see Methods and SI text). The ratios obtained with both approaches were 0.124 (95% CI 0.117-0.131) and 0.126 (95% CI 0.115-0.137), respectively. The two independent estimates are strikingly consistent with each other and robust to small changes in the methodology (SI text). Accordingly, we took the average $d/l = 0.125$ as the intrinsic duplication/loss ratio. This value quantifies the intrinsic bias towards 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. (3) 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 coefficients 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, 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 at 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, (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 anti-parasite defense (Mann-Whitney test, $p < 10^{-7}$). Thus, the drug resistance machinery is close to neutral whereas the anti-parasite 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 approximately 0.01 to 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-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 ~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 ~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 ~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, ~12% of the lineage-specific families of transposons show evidence of recent expansions, as indicated by effective duplication/loss ratios greater than one, whereas the fraction of such families drops below 4% in other functional categories (Fig. 2A, 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. 2A, 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 *X. citri*.

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, $\ll k \gg$, becomes $\ll k \gg = \int_0^\infty dt \ \sum_k k \, p_k(t) \ \phi \, e^{-\phi t}$. 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 text 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, see 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 to 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 text). 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 text). The same qualitative correlations were detected, i.e h/l positively correlates with N_e, whereas d/l shows no correlation. However, the association between h/l and N_e becomes non-significant 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 not only result 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 non-synonymous 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 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

298 than predicted from the overall trend (Fig. 4), suggesting that the effective neutrality of such genes results
299 from the heterogeneity of environmental conditions. Among the gene categories with low selection
300 coefficients, the *dN*/*dS* values of transposons, prophages and gene families with poorly characterized
301 functions are significantly greater than expected from the general trend, which is consistent with the
302 notion that these genes provide little or no benefit to the cells that harbor them.

303

304 **Gene dynamics and microbial life styles**

305 In an effort to clarify the biological underpinnings of the gene dynamics, we compared the effective
306 duplication to loss ratios in microbes with three life styles: free-living, facultative host-associated and
307 obligate intracellular parasite (Fig. 5). In the first two groups, d/l_e drops from essential functional
308 categories to non-essential categories and genetic parasites, with significantly higher values in free-living
309 microbes than in facultative host-associated bacteria. Obligate intracellular parasites have remarkably low
310 d/l_e values, as could be expected from their strong genomic degeneration. Notably, genetic parasites and
311 genes from the defense category show the highest d/l_e among the genes of intracellular parasites, although
312 due to the small number of intracellular parasites in our dataset (only 3 ATGCs, with most genetic
313 parasites restricted to the ATGC044 encompassing *Rickettsia*), this result must be taken with caution. We
314 estimated the selection coefficients for free-living and facultative host-associated microbes, under the
315 assumption that the intrinsic d/l is universally the same across the microbial diversity. The significant
316 difference in d/l_e between the two lifestyles translates into consistently higher s values for most functional
317 categories of genes in free-living microbes (SI Fig. S2). Thus, the beneficial effects of most genes appear
318 to be significantly greater in free-living compared to facultative host-associated bacteria, and in both these
319 categories of microbes, selection for gene retention is dramatically stronger than it is in obligate,
320 intracellular parasites.

321

322

323 **Discussion**

324 Multiple variants of the duplication-transfer-loss model and related multi-type branching processes have
325 been widely used to study the evolution of gene copy numbers (24, 25, 28, 49), especially in the context
326 of transposons and other genetic parasites (22, 23, 26, 27, 50). To make the models tractable, most studies
327 make simplifying assumptions, such as stationary state, absence of duplication or lack of selection, and
328 obtain the model parameters from the copy number distributions observed in large genomic datasets,

329 relying on the assumption that model parameters are homogeneous across taxa. Here we derived an exact
330 solution for the time-dependent duplication-transfer-loss model with additive selection and found that, in
331 general, it is impossible to distinguish neutral and costly elements solely based on the copy number
332 distributions. This is the case because the effects of selection and loss bias blend into a composite
333 parameter that is equivalent to an effective loss bias in a neutral scenario. Using the solution of the
334 complete model, we investigated the copy number dynamics of a large number of gene families in groups
335 of related genomes, without the need to assume homogeneity of the HGT, duplication and loss rates
336 across taxa (8). We then used the expression that relates the parameter values under selection with their
337 neutral equivalents to estimate the selection coefficients for different classes of genes.

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

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 lisogenization (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

394 selection against high copy numbers, present fewer degenerated copies than lower cost elements, such as
395 transposons.

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

407 Long-term transposon dynamics is well described by a model that combines long phases of decay, during
408 which transposons behave as inactive genetic material, punctuated by small proliferation bursts that
409 produce on average 4 new copies. Despite the simplicity of this model, it captures, at least qualitatively,
410 the heterogeneity of transposition rates among transposon families (65) and environmental conditions (66,
411 67). Unlike large expansions, which are rare events typically associated with ecological transitions
412 affecting the entire genome (68-71), small bursts occur frequently and affect a sizable fraction of
413 transposon families. Some well-known instances of large transposon expansions become apparent in our
414 analysis that identified taxa with unusually high burst rates, such as *Xanthomonas*, *Burkholderia* and
415 *Francisella*, in accord with previous observations (70, 71). In most other taxa, transposon decay is the
416 dominant process, which is the expected trend, given that transposition is tightly regulated and a large
417 fraction of transposon copies are inactive (72, 73). The small fitness cost of transposons in the decay
418 phase is also consistent with a non-proliferative scenario, where the fitness effect is reduced to the
419 energetic cost of replication and expression (21). Due to the rapidity of bursts, our methodology cannot be
420 used to assess the cost of a transposon during the burst phase. Because active transposons likely impose a
421 larger burden on the host (74), variation in burst sizes is likely to reflect differences in the intensity of
422 selection and the duration of proliferative episodes.

423 Apart from the transposons, the only notable case of burst-driven dynamics corresponds to genes from the
424 defense category in *Sulfolobus*. A closer inspection of this group reveals multiple instances of
425 duplications, gains and losses of CRISPR-Cas systems as also observed previously (75). In the case of
426 prophages, the low burst rate is likely to reflect genuine lack of bursts or our inability to detect them due

427 to the dominant, selection-driven fast decay dynamics. Indeed, given the fitness cost that we estimated for
428 prophages, a burst of prophages would decay almost three times faster than a burst of transposons of
429 similar size.

430 The effective size of microbial populations positively correlates with the genome size, which led to the
431 hypothesis that the genome dynamics is dominated by selection acting to maintain slightly beneficial
432 genes (14, 16). In the present analysis, when gene families from all functional categories are pooled, the
433 median fitness contribution per gene is $N_e s$ ~0.1, which provides independent support for this weak
434 selection-driven concept of microbial genome evolution. In that framework, the fact that genetic parasites
435 are more abundant in large genomes, as reported previously (76-78) and confirmed by our data,
436 seemingly raises a paradox: the same genomes where selection works more efficiently to maintain
437 beneficial genes also harbor more parasites. A possible solution comes from our observation that the HGT
438 to loss ratio (where the HGT rate is measured per genome and the loss rate is measured per gene) grows
439 with the genome size. Such behavior, which had been already noted for transposons (23) and agrees with
440 the recently derived genome-average scaling law (14), is likely to result, at least in part, from larger
441 genomes providing more non-essential regions where a parasite can integrate without incurring major
442 costs to the cell. Alternatively or additionally, the observed dependence could emerge if duplication and
443 loss rates per gene decreased with genome size, while the HGT rate remains constant. Indeed, an inverse
444 correlation between the genome size and the duplication and loss rates could be expected as long as
445 mutation rates appear to have evolved to lower values in populations with larger N_e (41, 79).

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

455 Counterintuitive as this might be, we show that anti-parasite defense systems are generally deleterious for
456 microbes, roughly to the same extent as mobile elements. These results are compatible with the
457 previously observed highly dynamic evolution of such systems that are kept by microbes either when they
458 are essential to counteract aggressive parasites or due to their own selfish and addictive properties. These
459 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 $G(z,t) = \sum_{k=0}^{\infty} z^k\, n_k(t)$. 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 $\rho = \frac{d}{l}$, $\beta = \frac{h}{l}$, and $\alpha = 1 + \rho - \frac{s}{l}$. This equation can be solved for any initial condition by applying the method of characteristics (SI text). 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 ATCG database that clusters genomes from bacteria and archaea into closely related groups (33). We analyzed 35 of the largest ATGCs (34 bacterial and one 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 post-processed to obtain ATGC-COG-specific rates as described in (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, respectively, as the median d/l_e and the sum of h/l_e 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 non-uniform ATGC sizes. To minimize inference artifacts associated to small families, only those families encompassing at least 5 ATGC-COGs from at least 3 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 $median \pm 1.7\ (1.25\ IQR/1.35\sqrt{N})$, 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 the SI text. 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 $LogLkh_{atgc} = n_{atgc} \log\left(1 - e^{-\phi t_{atgc}}\right) - \left(N_{atgc} - n_{atgc}\right) \phi\ t_{atgc}$, where t_{atgc} is the depth of the ATGC tree in units of loss events (SI text). 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 package (84) was used to obtain the *dN*/*dS* for each pair of

521 sequences in the ATGC-COG. The conditions $0.01 < dN < 3$ and $0.01 < dS < 3$ were used to select
522 informative gene pairs. The representative dN/dS for the ATGC-COG was obtained as the median dN/dS
523 among the informative pairs. In the next step, ATGC-COGs from the same ATGC that belong to the same
524 functional category were pooled and the median of their dN/dS was taken as the representative dN/dS.
525 To account for ATGC-related effects, the dN/dS values of all categories within an ATGC were converted
526 into ranks. The null hypothesis that all categories are equal in terms of their dN/dS was rejected by a
527 Skillings-Mack test ($t = 939.7$, d.f. = 22, $p < 10^{-20}$). To identify which categories significantly deviate from
528 the null hypothesis, the mean rank of each category was compared to the theoretical 95% CI for the
529 mean of 35 samples taken from a discrete uniform distribution in the interval from 1 to 23.

530

531

532

1. Koonin EV (2011) *The Logic of Chance: The Nature and Origin of Biological Evolution* (FT press, Upper Saddle River, NJ).
2. Lynch M (2007) *The origins of genome archiecture* (Sinauer Associates, Sunderland, MA).
3. Koonin EV & Wolf YI (2012) Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front Cell Infect Microbiol* 2.
4. Moran NA & Bennett GM (2014) The tiniest tiny genomes. *Annual review of microbiology* 68:195-215.
5. Han K*, et al.* (2013) Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. *Scientific reports* 3:2101.
6. Koonin EV (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* 1:127-136.
7. Tettelin H, Riley D, Cattuto C, & Medini D (2008) Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol* 11(5):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 biology* 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(21):6688-6719.
10. Wolf YI, Makarova KS, Lobkovsky AE, & Koonin EV (2016) Two fundamentally different classes of microbial genes. *Nature microbiology* 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(9):722-732.
12. Lobkovsky AE, Wolf YI, & Koonin EV (2013) Gene frequency distributions reject a neutral model of genome evolution. *Genome biology and evolution* 5(1):233-242.
13. Kuo CH & Ochman H (2009) Deletional bias across the three domains of life. *Genome biology and evolution* 1:145-152.
14. Sela I, Wolf YI, & Koonin EV (2016) Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* 113(41):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. *Journal of bacteriology* 191(1):65-73.

16. Kuo CH, Moran NA, & Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome research* 19(8):1450-1454.

17. Kuo CH & Ochman H (2010) The extinction dynamics of bacterial pseudogenes. *PLoS genetics* 6(8).

18. Lee M-C & Marx CJ (2012) Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet.* 8(5):e1002651.

19. Koskiniemi S, Sun S, Berg OG, & Andersson DI (2012) Selection-driven gene loss in bacteria. *PLoS Genet.* 8(6):e1002787.

20. Iranzo J, Puigbo P, Lobkovsky AE, Wolf YI, & Koonin EV (2016) Inevitability of Genetic Parasites. *Genome biology and evolution* 8(9):2856-2869.

21. Lynch M & Marinov GK (2015) The bioenergetic costs of a gene. *Proceedings of the National Academy of Sciences of the United States of America* 112(51):15690-15695.

22. Bichsel M, Barbour AD, & Wagner A (2013) Estimating the fitness effect of an insertion sequence. *Journal of mathematical biology* 66(1-2):95-114.

23. Iranzo J, Gomez MJ, Lopez de Saro FJ, & Manrubia S (2014) Large-scale genomic analysis suggests a neutral punctuated dynamics of transposable elements in bacterial genomes. *PLoS computational biology* 10(6):e1003680.

24. Csuros M (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26(15):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 evolutionary biology* 4.

26. van Passel MWJ, Nijveen H, & Wahl LM (2014) Birth, Death, and Diversification of Mobile Promoters in Prokaryotes. *Genetics* 197(1):291-299.

27. Basten CJ & Moody ME (1991) A Branching-Process Model for the Evolution of Transposable Elements Incorporating Selection. *Journal of mathematical biology* 29(8):743-761.

28. Huynen MA & van Nimwegen E (1998) The frequency distribution of gene family sizes in complete genomes. *Molecular biology and evolution* 15(5):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 evolutionary biology* 2(1):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(Database issue):D448-454.
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(D1):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(Database issue):D261-269.
35. Siew N & Fischer D (2003) Unravelling the ORFan Puzzle. *Comp Funct Genomics* 4(4):432-441.
36. Yu G & Stoltzfus A (2012) Population diversity of ORFan genes in Escherichia coli. *Genome biology and evolution* 4(11):1176-1187.
37. Nilsson AI*, et al.* (2005) Bacterial genome size reduction by experimental evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102(34):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(3):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. *Molecular biology and evolution* 34(1):93-109.
40. Long H*, et al.* (2015) Background Mutational Features of the Radiation-Resistant Bacterium Deinococcus radiodurans. *Molecular biology and evolution* 32(9):2383-2392.
41. Sung W*, et al.* (2016) Evolution of the Insertion-Deletion Mutation Rate Across the Tree of Life. *G3* 6(8):2583-2591.
42. Sung W*, et al.* (2015) Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation-Accumulation Experiments. *Molecular biology and evolution* 32(7):1672-1683.
43. Sung W, Ackerman MS, Miller SF, Doak TG, & Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. *Proceedings of the National Academy of Sciences of the United States of America* 109(45):18488-18492.

627 44. Sung W*, et al.* (2012) Extraordinary genome stability in the ciliate Paramecium tetraurelia.
628 *Proceedings of the National Academy of Sciences of the United States of America* 109(47):19339-
629 19344.
630 45. Dettman JR, Sztepanacz JL, & Kassen R (2016) The properties of spontaneous mutations in the
631 opportunistic pathogen Pseudomonas aeruginosa. *BMC genomics* 17:27.
632 46. Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annual review*
633 *of microbiology* 60:327-349.
634 47. Lynch M & Conery JS (2003) The origins of genome complexity. *Science* 302(5649):1401-1404.
635 48. Krylov DM, Wolf YI, Rogozin IB, & Koonin EV (2003) Gene loss, protein sequence divergence,
636 gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution.
637 *Genome research* 13(10):2229-2235.
638 49. Novozhilov AS, Karev GP, & Koonin EV (2006) Biological applications of the theory of birth-and-
639 death processes. *Brief Bioinform* 7(1):70-85.
640 50. Moody ME (1988) A Branching-Process Model for the Evolution of Transposable Elements.
641 *Journal of mathematical biology* 26(3):347-357.
642 51. Veitia RA & Potier MC (2015) Gene dosage imbalances: action, reaction, and models. *Trends in*
643 *biochemical sciences* 40(6):309-317.
644 52. Axelsen JB, Yan KK, & Maslov S (2007) Parameters of proteome evolution from histograms of
645 amino-acid sequence identities of paralogous proteins. *Biology direct* 2:32.
646 53. Innan H & Kondrashov F (2010) The evolution of gene duplications: classifying and distinguishing
647 between models. *Nature reviews. Genetics* 11(2):97-108.
648 54. Jalasvuori M & Koonin EV (2015) Classification of prokaryotic genetic replicators: between
649 selfishness and altruism. *Annals of the New York Academy of Sciences* 1341(1):96-105.
650 55. Cerveau N, Leclercq S, Leroy E, Bouchon D, & Cordaux R (2011) Short- and long-term
651 evolutionary dynamics of bacterial insertion sequences: insights from Wolbachia
652 endosymbionts. *Genome biology and evolution* 3:1175-1186.
653 56. Nelson WC, Wollerman L, Bhaya D, & Heidelberg JF (2011) Analysis of insertion sequences in
654 thermophilic cyanobacteria: exploring the mechanisms of establishing, maintaining, and
655 withstanding high insertion sequence abundance. *Applied and environmental microbiology*
656 77(15):5458-5466.
657 57. Brugger K*, et al.* (2002) Mobile elements in archaeal genomes. *FEMS microbiology letters*
658 206(2):131-141.

659 58. Van Melderen L (2010) Toxin-antitoxin systems: why so many, what for? *Curr Opin Microbiol*
660 13(6):781-785.
661 59. Van Melderen L & Saavedra De Bast M (2009) Bacterial toxin-antitoxin systems: more than
662 selfish entities? *PLoS genetics* 5(3):e1000437.
663 60. Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and
664 their impact on genome evolution. *Nucleic Acids Res* 29(18):3742-3756.
665 61. Furuta Y & Kobayashi I (2011) Restriction-modification systems as mobile epigenetic elements.
666 *Bacterial Integrative Mobile Genetic Elements*, eds Roberts AP & Mullany P (Landes Bioscience,
667 Austin, TX).
668 62. Weinberger AD, Wolf YI, Lobkovsky AE, Gilmore MS, & Koonin EV (2012) Viral diversity threshold
669 for adaptive immunity in prokaryotes. *MBio* 3(6):e00456-00412.
670 63. Koonin EV & Zhang F (2017) Coupling immunity and programmed cell suicide in prokaryotes:
671 Life-or-death choices. *Bioessays* 39(1):1-9.
672 64. Iranzo J, Lobkovsky AE, Wolf YI, & Koonin EV (2015) Immunity, suicide or both? Ecological
673 determinants for the combined evolution of anti-pathogen defense systems. *BMC evolutionary*
674 *biology* 15:43.
675 65. Sousa A, Bourgard C, Wahl LM, & Gordo I (2013) Rates of transposition in Escherichia coli. *Biol.*
676 *Lett.* 9(6):20130838.
677 66. Ohtsubo Y, Genka H, Komatsu H, Nagata Y, & Tsuda M (2005) High-temperature-induced
678 transposition of insertion elements in burkholderia multivorans ATCC 17616. *Applied and*
679 *environmental microbiology* 71(4):1822-1828.
680 67. Naas T, Blot M, Fitch WM, & Arber W (1994) Insertion sequence-related genetic variation in
681 resting Escherichia coli K-12. *Genetics* 136(3):721-730.
682 68. Beare PA*, et al.* (2009) Comparative genomics reveal extensive transposon-mediated genomic
683 plasticity and diversity among potential effector proteins within the genus Coxiella. *Infection*
684 *and immunity* 77(2):642-656.
685 69. Moran NA & Plague GR (2004) Genomic changes following host restriction in bacteria. *Current*
686 *opinion in genetics & development* 14(6):627-633.
687 70. Mira A, Pushker R, & Rodriguez-Valera F (2006) The Neolithic revolution of bacterial genomes.
688 *Trends in microbiology* 14(5):200-206.
689 71. Rohmer L*, et al.* (2007) Comparison of Francisella tularensis genomes reveals evolutionary
690 events associated with the emergence of human pathogenic strains. *Genome biology* 8(6):R102.

691 72. Nagy Z & Chandler M (2004) Regulation of transposition in bacteria. *Research in microbiology*
692 155(5):387-398.
693 73. Filee J, Siguier P, & Chandler M (2007) Insertion sequence diversity in archaea. *Microbiology and*
694 *molecular biology reviews : MMBR* 71(1):121-157.
695 74. Elena SF, Ekunwe L, Hajela N, Oden SA, & Lenski RE (1998) Distribution of fitness effects caused
696 by random insertion mutations in Escherichia coli. *Genetica* 102-103(1-6):349-358.
697 75. Garrett RA*, et al.* (2011) CRISPR-based immune systems of the Sulfolobales: complexity and
698 diversity. *Biochem Soc Trans* 39(1):51-57.
699 76. Zhou F, Olman V, & Xu Y (2008) Insertion Sequences show diverse recent activities in
700 Cyanobacteria and Archaea. *BMC genomics* 9:36.
701 77. Touchon M & Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic
702 genomes. *Molecular biology and evolution* 24(4):969-981.
703 78. Touchon M, Bernheim A, & Rocha EP (2016) Genetic and life-history traits associated with the
704 distribution of prophages in bacteria. *The ISME journal* 10(11):2744-2754.
705 79. Lynch M*, et al.* (2016) Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev.*
706 *Genet.* 17(11):704-714.
707 80. Kristensen DM*, et al.* (2010) A low-polynomial algorithm for assembling clusters of orthologous
708 groups from intergenomic symmetric best matches. *Bioinformatics* 26(12):1481-1487.
709 81. Csuros M (2006) On the estimation of intron evolution. *PLoS computational biology* 2(7):e84;
710 author reply e83.
711 82. McGill R, Tukey JW, & Larsen WA (1978) Variations of box plots. *Am. Stat.* 32(1):12-16.
712 83. Hudson DJ (1971) Interval Estimation from Likelihood Function. *J Roy Stat Soc B* 33(2):256-262.
713 84. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and*
714 *evolution* 24(8):1586-1591.
715 85. Kryazhimskiy S & Plotkin JB (2008) The population genetics of dN/dS. *PLoS genetics*
716 4(12):e1000304.

717

718

FIGURE LEGENDS

Figure 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 (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, post-translational 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; Pl, conjugative plasmid; 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.

Figure 2: Frequency and distribution of proliferation bursts in different functional categories of genes. A: Orange (left axis), 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), 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.

Figure 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; non-significant (n.s.) p-values are greater than 0.2.

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

760

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

765

766 **TABLES**

767 **Table 1:** Contributions of selection and the duplication/loss ratio to the evolution of different functional
768 categories of genes and mobile elements

769

	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 - post-translational 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) – Anti-pathogen 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 fitness cost (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.

FIGURE 1

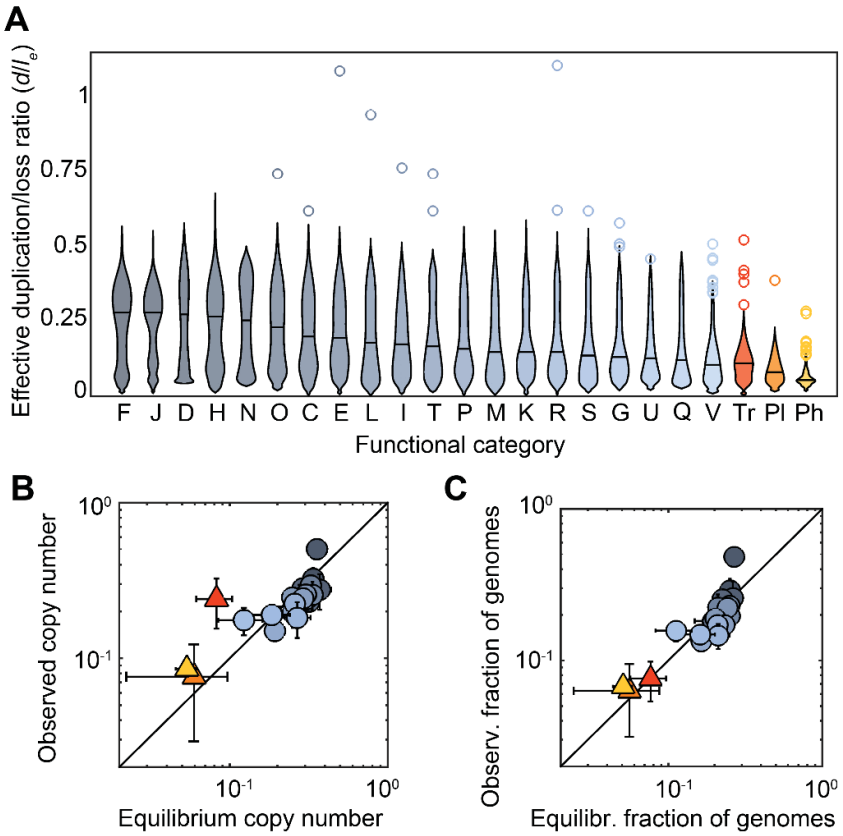

FIGURE 2

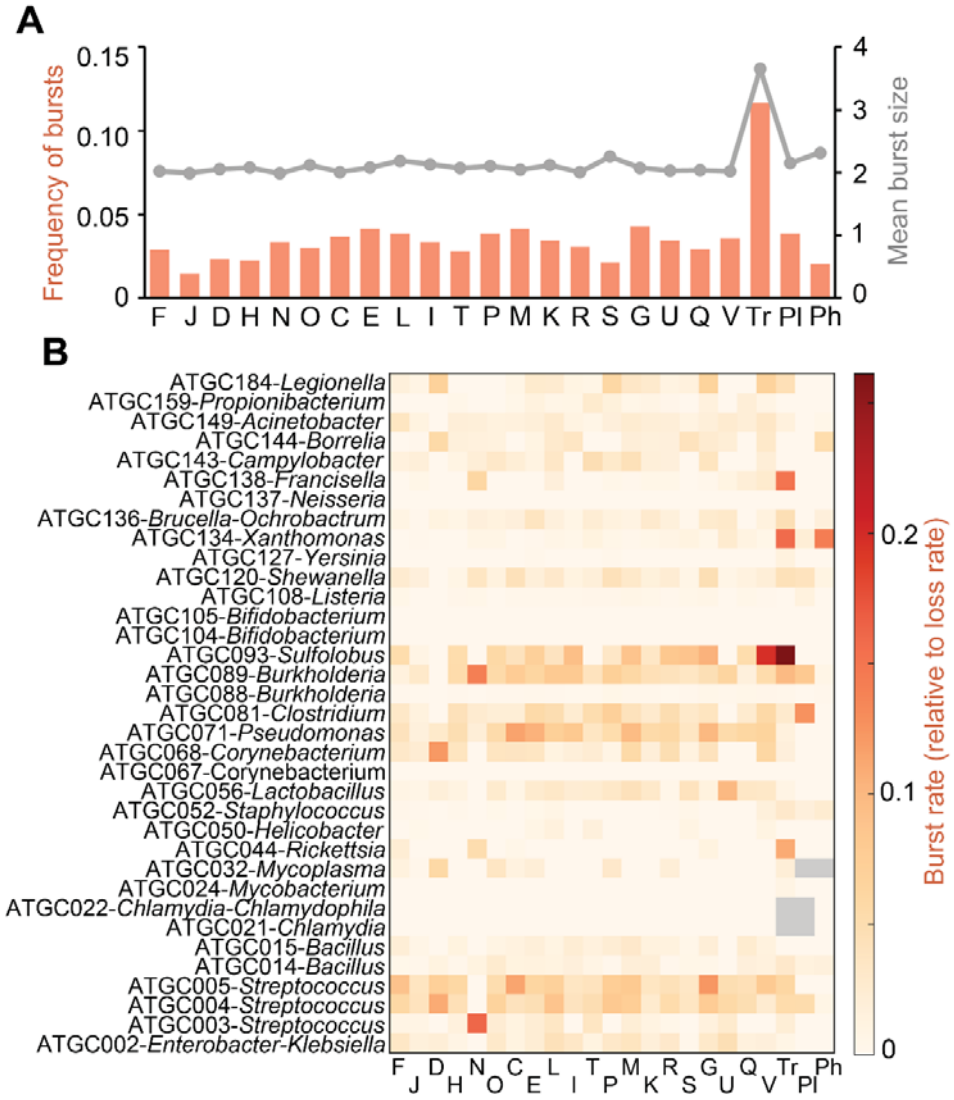

FIGURE 3

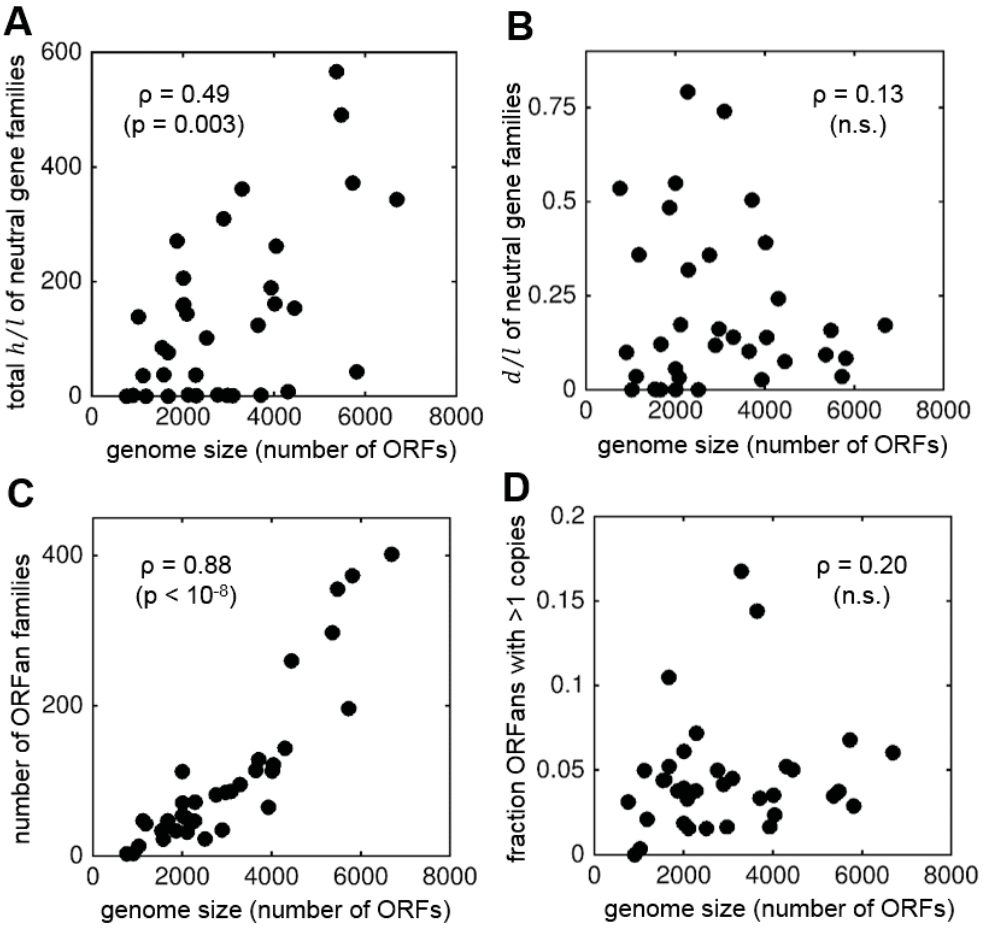

FIGURE 4

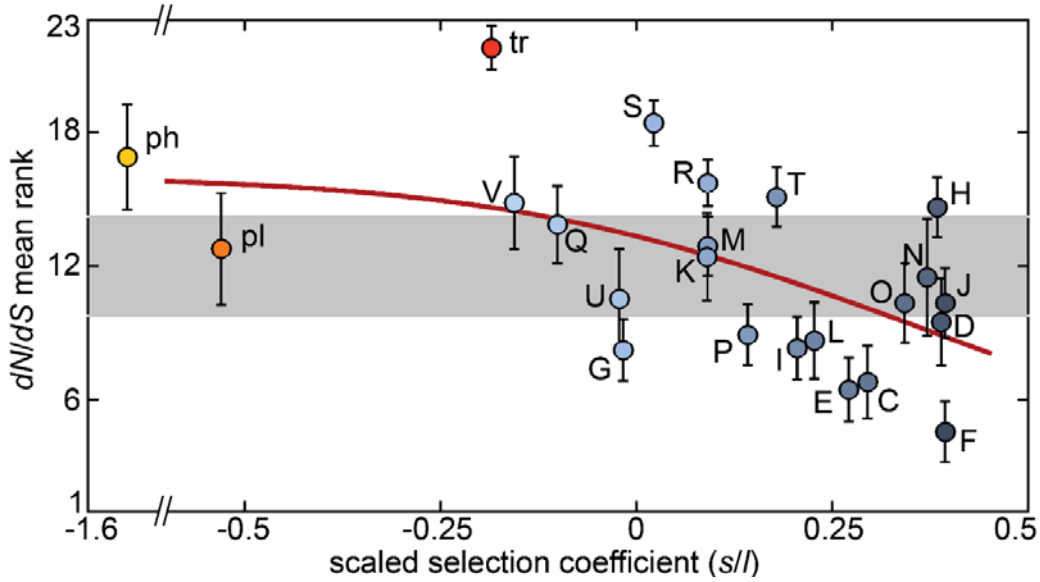

FIGURE 5