Selection and Genome Plasticity as the Key Factors in the Evolution of Bacteria

Itamar Sela,[*] Yuri I. Wolf, and Eugene V. Koonin[†]
National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, Maryland 20894, USA

(Received 10 December 2018; revised manuscript received 6 June 2019; published 5 August 2019)

In prokaryotes, the number of genes in different functional classes shows apparent universal scaling with the total number of genes that can be approximated by a power law, with a sublinear, near-linear, or superlinear scaling exponent. These dependences are gene class specific but hold across the entire diversity of bacteria and archaea. Several models have been proposed to explain these universal scaling laws, primarily based on the specifics of the respective biological functions. However, a population-genetic theory of universal scaling is lacking. We employ a simple mathematical model for prokaryotic genome evolution, which, together with the analysis of 34 clusters of closely related bacterial genomes, allows us to identify the underlying factors that govern the evolution of the genome content. Evolution of the gene content is dominated by two functional class-specific parameters: selection coefficient and genome plasticity. The selection coefficient quantifies the fitness cost associated with deletion of a gene in a given functional class or the advantage of successful incorporation of an additional gene. Genome plasticity reflects both the availability of the genes of a given class in the external gene pool that is accessible to the evolving population and the ability of microbes to accommodate these genes in the short term, that is, the class-specific horizontal gene transfer barrier. The selection coefficient determines the gene loss rate, whereas genome plasticity is the principal determinant of the gene gain rate.

DOI: 10.1103/PhysRevX.9.031018 Subject Areas: Biological Physics, Interdisciplinary Physics

I. INTRODUCTION

Comparative analyses of prokaryotic genomes show that the number of genes in different functional classes scales differentially with the genome size [1–6]. The scaling laws are robust under various statistical tests [5], across different databases, and for different gene classifications [2–6]. In the seminal analysis of scaling, van Nimwegen fitted the scaling to a power law of the form [2]

$$x_1 = \eta x^{\gamma}, \tag{1}$$

where x_1 denotes the number of genes that belong to a specific functional class, and x is the total number of genes. Power laws are the simplest functions that give good fits to the gene scaling data [2,5]. Analysis of the scaling exponents γ has shown that such exponents are (nearly) universal for each functional class across a broad range of microbes (notwithstanding some debate on the validity of the exact universality [5,7]), suggesting that differences in scaling reflect important not yet understood features of cellular organization and its evolution. In attempts to explain the empirical observation that power-law scaling is a good fit to the genomic data, several theoretical models have been proposed, as outlined below.

In the first, now classic analysis of scaling laws, van Nimwegen grouped the functional classes of genes along three integer exponents 0,1,2, arguing that deviations from the integers, as demonstrated in the dataset analyzed here (Fig. 1 and Table I), most likely reflected gene classification ambiguities [2]. The gene classes with the 0 exponent include information processing systems (translation, basal transcription, and replication), those with the exponent of 1 are primarily genes for metabolic enzymes and transporters, whereas those with the exponent of 2 encode various regulatory proteins. The essential information processing systems are universally conserved and remain nearly the same in all microbes regardless of genome size; metabolic networks expand proportionally to the genome growth, and the complexity of regulatory circuits increases quadratically with the total number of genes (i.e., linearly with the number of potential interactions between gene products). To put this conceptual thinking on quantitative ground, the toolbox model has been proposed to explain the quadratic

[*] itamar.sela@nih.gov
[†] koonin@ncbi.nlm.nih.gov

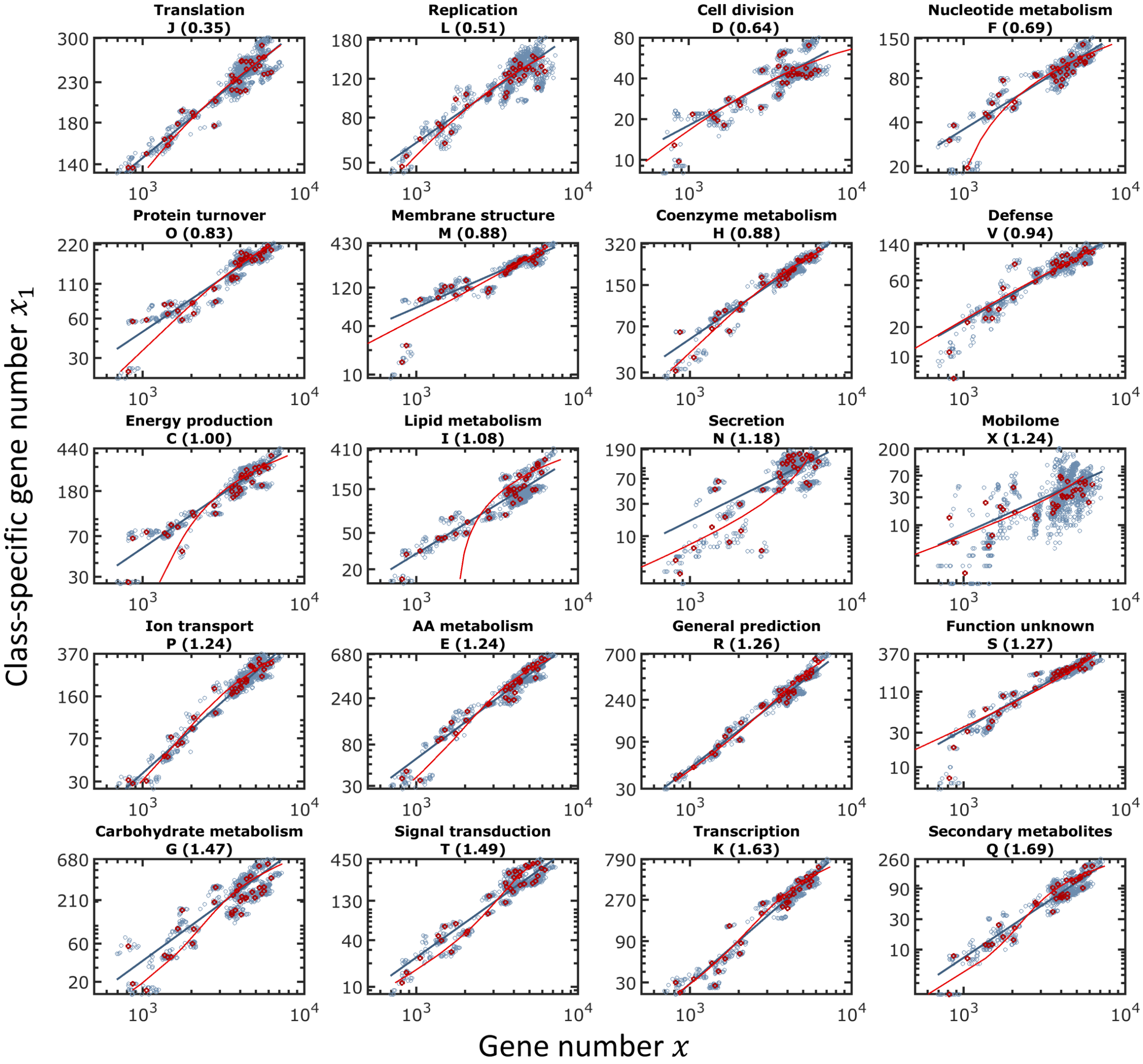


FIG. 1. The scaling laws for all functional classes of the COGs. The number of genes in each functional class scales with the total number of genes, and the scaling exponents substantially differ between the classes. The number of genes in a given COG category is plotted against the total number of genes, together with a power-law fit and model fit, as given by Eq. (2). Each gray point represents one genome from the analyzed set of 1490 genomes. The scaling is fitted to a power law which is indicated by a solid gray line. The fitted scaling exponent is indicated in parentheses. Red points correspond to the mean values for each ATGC in the dataset, and the model fit of Eq. (2) is shown by the solid red line.

scaling, whereby the number of regulators grows faster than the number of metabolic enzymes thanks to the frequent reuse of the latter enzymes in new pathways [8,9]. Subsequently, the toolbox model has been extended to assume that the gene composition of prokaryotic genomes is determined by selection for fixed proportions of genes from different functional classes; under this assumption, the model recapitulates both the scaling laws and the distribution of gene family sizes within each functional class [10]. The linear scaling was also obtained under a theoretical model with two classes of genes, internal (housekeeping) and external (ecological), and a reproductive rate set to be independent of the genome size [11]. Finally, given that prokaryotic genome evolution is dominated by extensive gene loss and horizontal gene transfer (HGT) [12–16], it has been hypothesized that the universal exponents are determined by distinct gene gain and loss rates for different classes of genes and reflect the innovation potential of these classes [17]. Clearly, regulatory genes have the highest innovation potential, whereas information processing systems have next to none.

Although the power laws provide good fits to the genomic data, the origin of the observed scaling remains obscure. Given that genome sizes barely span 2 orders of magnitude (Fig. 1), the power-law fits should be treated as approximations rather than firmly established quantitative laws. More importantly, although the models surveyed above account for the power-law fit to the genomic data,

TABLE I. Scaling, selection, and plasticity in different functional classes of microbial genes.

Class	Functions	Scaling exponent γ	ΔS_1 slope $-q$	Average selection coefficient $\langle \Delta S_1 \rangle$	Average plasticity $\langle p_1 \rangle$	Plasticity slope b
J	Translation	0.35	-1.10×10^{-2}	2.68	0.005	1.54×10^{-5}
L	Replication and repair	0.51	-1.35×10^{-3}	0.98	0.013	-9.18×10^{-5}
D	Cell division	0.64	-1.58×10^{-5}	1.83	0.002	-3.21×10^{-5}
F	Nucleotide metabolism and transport	0.69	-1.75×10^{-2}	2.15	0.003	3.65×10^{-5}
O	Post-translational modification, protein turnover, and chaperone functions	0.83	-5.42×10^{-3}	1.38	0.010	4.88×10^{-5}
M	Membrane and cell wall structure and biogenesis	0.88	-1.05×10^{-3}	0.65	0.029	4.57×10^{-5}
H	Coenzyme metabolism	0.88	-4.68×10^{-3}	1.51	0.011	4.62×10^{-5}
V	Defense	0.94	-3.72×10^{-7}	−0.44	0.034	-6.68×10^{-5}
C	Energy production and conversion	1.00	-4.52×10^{-3}	1.19	0.017	8.92×10^{-5}
I	Lipid metabolism	1.08	-4.86×10^{-3}	0.92	0.016	1.24×10^{-4}
N	Secretion and motility	1.18	-7.13×10^{-8}	0.27	0.014	1.07×10^{-4}
X	Mobilome: prophages, transposons	1.24	-2.06×10^{-4}	−4.76	1.021	1.12×10^{-2}
P	Inorganic ion transport and metabolism	1.24	-3.76×10^{-3}	0.75	0.026	1.15×10^{-4}
E	Amino acid metabolism and transport	1.24	-1.90×10^{-3}	1.12	0.028	8.15×10^{-5}
R	General functional prediction only	1.26	-1.05×10^{-3}	0.37	0.051	1.15×10^{-4}
S	Function unknown	1.27	-5.92×10^{-7}	0.55	0.025	3.35×10^{-5}
G	Carbohydrate metabolism and transport	1.47	-1.86×10^{-3}	0.46	0.041	1.65×10^{-4}
T	Signal transduction	1.49	-1.28×10^{-3}	0.57	0.030	1.25×10^{-4}
K	Transcription	1.63	-1.67×10^{-3}	0.27	0.058	1.81×10^{-4}
Q	Biosynthesis, transport, and catabolism of secondary metabolites	1.69	-5.77×10^{-3}	0.06	0.021	2.54×10^{-4}

they stop short of a general theory of genome evolution rooted in population genetics that would yield power laws or even account for the observed scaling.

Here, we analyze a simple population genetics model for prokaryotic genome evolution [18]. Genome evolution is modeled as a stochastic process of gene gains and losses, and we formulate an explicit model for the gene gain and loss rates within the theory of population genetics. In previous studies, this modeling framework was developed and utilized to analyze the evolution of prokaryotic genome size [18,19], i.e., the number of genes. Here, we present two substantial extensions to the model that enable us to analyze the scaling laws and extract the underlying evolutionary factors. First, within the same modeling framework, we analyze the evolution of distinct functional classes of genes. Second, we analyze, also in a class-specific manner, the divergent evolution of genome content which is measured by the number of orthologs shared by a pair of genomes [20]. The scaling we obtain under this model does not follow a power law and does not yield integer exponents. Extraction of model parameters from the genomic data engenders two major challenges. First, it is essential to extract independently the factors that dictate gain and loss rates. This is achieved by accounting for the divergence of genome content which depends on the loss rate only. Second, to infer the dependence of the evolutionary factors on the genome size, it is essential to compare different groups of microbes that evolve under specific local influences [19]. To filter out such influences, class-specific quantities are normalized by genomic means, and the ratios are used to extract model parameters.

The analyses presented here show that prokaryotic evolution is dominated by two underlying factors: selection coefficient and genome plasticity. While the selection coefficient is a standard quantity in population genetics, genome plasticity is an evolutionary factor that emerged from the analysis presented here, and it is the principle determinant of the gene gain rate. The class-specific genome plasticity reflects both the abundance of the genes of a given functional class in the external gene pool from which genes can be captured by the evolving microbial population, and the class-specific HGT barrier, i.e., the ability of genomes to absorb new genes from the given class. The HGT barrier not only decreases with the number of genes, but the reduction in the barrier height is class specific and dictates the scaling. To illustrate these findings, two representative genomes of different sizes are illustrated in Fig. 2 as collections of genes [Figs. 2(a) and 2(b)]. The reduction in the HGT-barrier height is modest in functional classes that scale sublinearly and comprise a similar fraction of the genome across different groups of microbes, largely, independent of the genome size [Fig. 2(c)]. In contrast, in the classes that scale superlinearly with the genome size, the reduction in the HGT-barrier height is dramatic, allowing for frequent acquisition and fast turnover rate of the respective genes. Thus, in this work, we present a simple population-genetic theory that uncovers the evolutionary factors underlying the observed universal

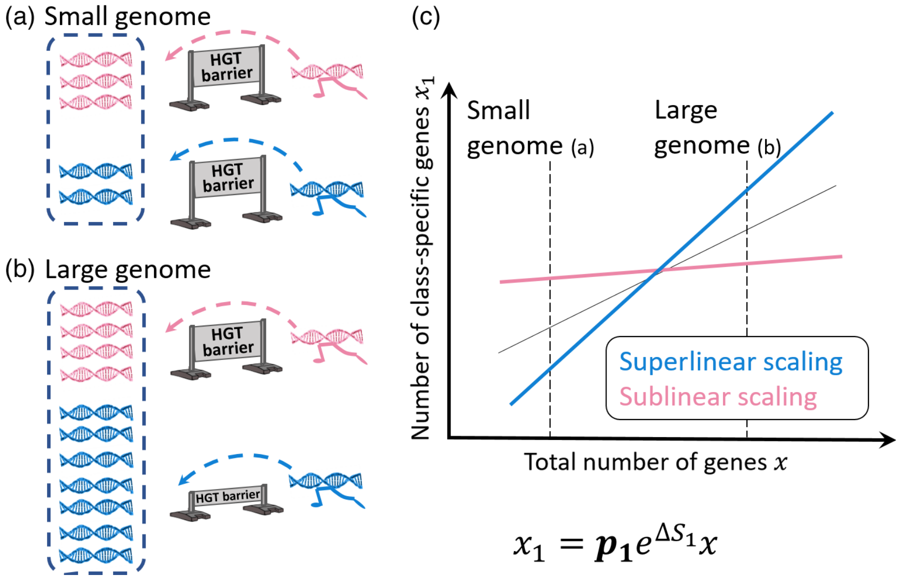


FIG. 2. Horizontal gene transfer barrier determines the scaling of class-specific genes. The gene-class-specific HGT barrier is high in small genomes (a) but substantially lower in large genomes (b). The dependence of the HGT barrier height on the genome size determines the scaling of the number of genes in each functional class (x_1) with the total number of genes x (c). For functional classes with a weak dependence of the HGT barrier height on the genome size (e.g., genes for translation system components), the scaling is sublinear. In contrast, for those classes that show a strong dependence such that the HGT barrier is substantially lowered in larger genomes, the scaling is superlinear (e.g., genes for transcription factors). This figure shows schematically how the HGT barrier determines that scaling. Under our model, the class-specific HGT barrier is reflected by the genome plasticity p_1.

scaling of gene functional classes with genome size in prokaryotes.

II. RESULTS

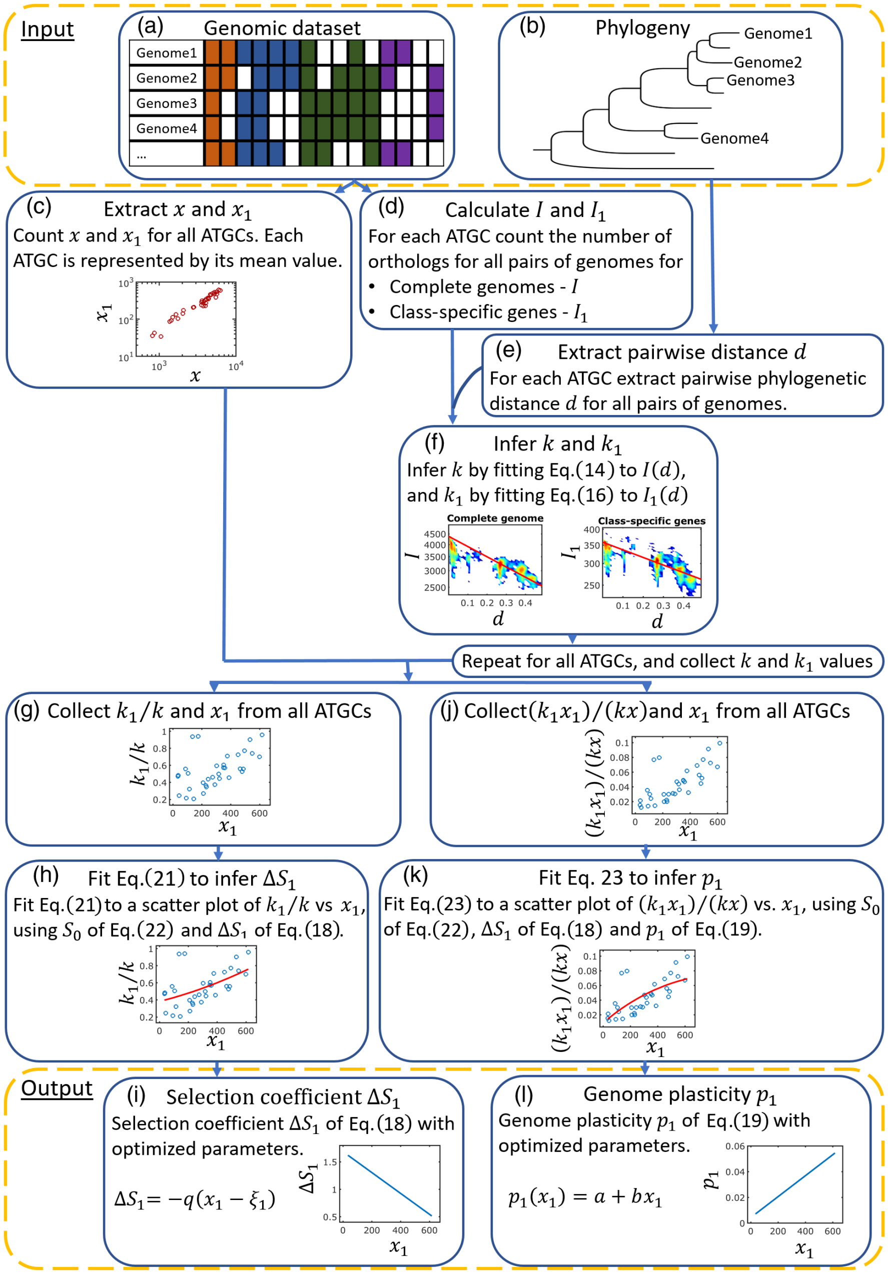
A. Genomic data

The dataset analyzed here consists of 34 clusters of bacterial genomes taken from the Alignable Tight Genomic Clusters (ATGC) database [21]. Each cluster contains ten or more genomes (see the Appendix A for details) and represents a sample of closely related genomes, such that the maximum evolutionary distance between pairs of genomes in each cluster is of the order of 0.1 over the set of core genes, in units of the mean number of substitutions per site. All genomes in each cluster are fully annotated, and for each genome cluster, genes are grouped into clusters of orthologs (ATGC COGs), such that each genome can be represented as an array indicating the presence or absence of ATGC COGs. In addition, all ATGC COGs are assigned to functional classes according to the functional classification of the Clusters of Orthologous Groups (COG) database [22]. Finally, phylogenetic trees for each ATGC are also available. A graphical representation of a single ATGC in the dataset is shown in Figs. 3(a) and 3(b).

B. The scaling laws

We analyze scaling with the genome size for 20 functional classes of genes from the database of COGs [22]. For each functional class of genes, a power-law fit is obtained and the scaling exponent is determined (Fig. 1 and Table I). Extraction of the scaling from the genomic dataset is depicted in Fig. 3(c), and the power-law fitting scheme is detailed in Appendix B. The scaling exponents range from 0.35 for translation genes (J COG category) to 1.69 for secondary biosynthesis genes (Q COG category) (Table I). It should be noted that the transcription category has an exponent of 1.63 (rather than the previously reported quadratic scaling), most likely because in the COG classification, it includes both basal transcription proteins that, in the previous analyses, show exponents close to 0 and transcription regulators with the apparent quadratic dependence on the total number of genes [6]. The observed values of gene-class-specific exponents show a broad range from sublinear for the essential universal information transmission genes to superlinear for more evolutionarily volatile genome components, such as regulators and secondary metabolism enzymes (Table I).

The robustness of the observed scaling exponents for different classes is tested by bootstrap analysis (Fig. 4; see Appendix C). Although for some of the functional classes of genes, the distribution of the bootstrap scaling exponents is wide [e.g., secretion and motility genes (N); Fig. 4], all classes can be confidently partitioned into those scaling sublinearly, near linearly, or superlinearly. This classification is important because it captures qualitatively different behaviors that are robust with respect to the collection of genomes analyzed. The key question we address is what are



(c) Extract x and x_1

Count x and x_1 for all ATGCs. Each ATGC is represented by its mean value.

(d) Calculate I and I_1

For each ATGC count the number of orthologs for all pairs of genomes for

- Complete genomes - I
- Class-specific genes - I_1

(e) Extract pairwise distance d

For each ATGC extract pairwise phylogenetic distance d for all pairs of genomes.

(f) Infer k and k_1

Infer k by fitting Eq.(14) to $I(d)$, and k_1 by fitting Eq.(16) to $I_1(d)$

Repeat for all ATGCs, and collect k and k_1 values

(g) Collect k_1/k and x_1 from all ATGCs

(j) Collect$(k_1x_1)/(kx)$and x_1 from all ATGCs

(h) Fit Eq.(21) to infer ΔS_1

Fit Eq.(21)to a scatter plot of k_1/k vs x_1, using S_0 of Eq.(22) and ΔS_1 of Eq.(18).

(k) Fit Eq. 23 to infer p_1

Fit Eq.(23) to a scatter plot of $(k_1x_1)/(kx)$ vs. x_1, using S_0 of Eq.(22), ΔS_1 of Eq.(18) and p_1 of Eq.(19).

(i) Selection coefficient ΔS_1

Selection coefficient ΔS_1 of Eq.(18) with optimized parameters.

$\Delta S_1 = -q(x_1 - \xi_1)$

(l) Genome plasticity p_1

Genome plasticity p_1 of Eq.(19) with optimized parameters.

$p_1(x_1) = a + bx_1$

FIG. 3. Graphical visualization of the analysis for a single functional class. This figure illustrates the stages of the model parameters inference scheme from the genomic data starting from the input of the genomic data (Sec. II A) to the output of the fitted model parameters of Eqs. (18) and (19). (a) Visualization of the dataset used as the model input. A representation of four genomes from a single cluster of genomes (ATGC) is shown. The genomic data can be represented as a table where each row corresponds to a genome and each column corresponds to a gene. Colored entries indicate presence of a gene in a genome, and different colors indicate different functional classes of genes. (b) The evolutionary relationships between the genomes are represented by a phylogenetic tree. (c) The total number of genes and the number of class-specific genes are extracted for each ATGC. Each ATGC is represented by a circle, where the x_1 and x values for each ATGC are taken as the mean values for all genomes in the ATGC. (d) For each ATGC, the number of common orthologs is counted for all genome pairs, both for all genes, and for functional classes of genes. (e) For all genome pairs, the evolutionary distance is inferred from the phylogenetic tree shown in (b). (f) Decay constants of genome intersections with evolutionary distance k and k_1 are inferred from the genomic data by fitting the model to observed decays of I or I_1 (model fits are shown by red lines). The genomic data are shown by a heat map [see the legend to Fig. 5(a) for details]. (g) The analysis described in (f) is performed for all ATGCs to capture the dependence of the k_1/k ratio on x_1. Each point in the scatter plot represents a single ATGC, where the representative x_1 value for an ATGC is taken as the mean value from all genomes in the cluster. (h) Inference of ΔS_1 by fitting Eq. (21) to the data of (g). The data are shown by circles, and the fitted analytical curve is shown by a red line. (i) Inferred ΔS_1. (j) The same as (g), but the ratio $(k_1x_1)/(kx)$. (k) Inference of p_1 by fitting Eq. (23) to the data of (j). The data are shown by circles, and the fitted analytical curve is shown by a red line. (l) Inferred p_1.

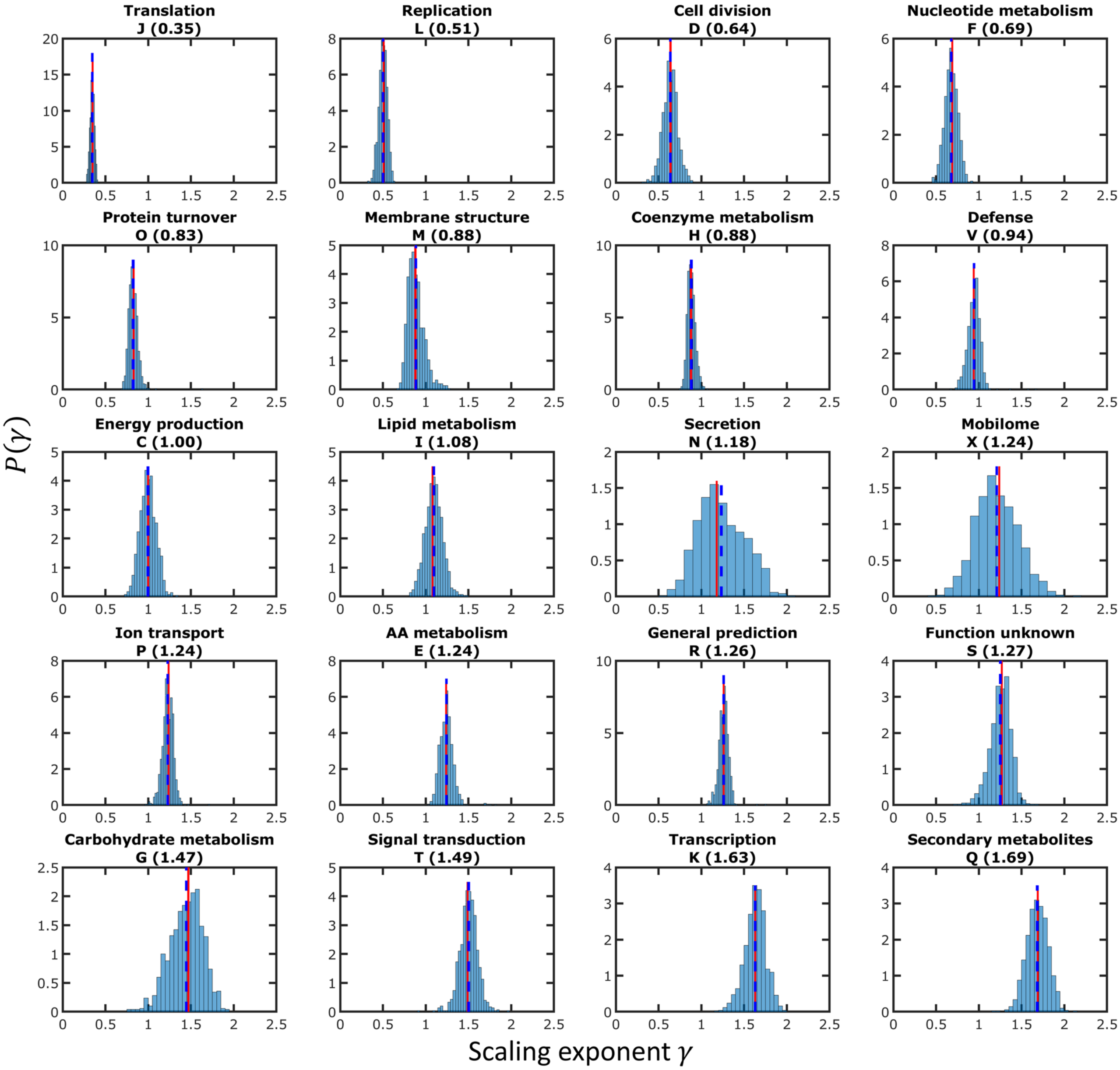


FIG. 4. Statistical support for scaling exponents calculated using bootstrap analysis. The distribution of fitted scaling exponents is shown for each class, for 1000 bootstrap samplings (see Appendix C). The mean of the distributions is indicated by a vertical dashed blue line, and the fitted scaling exponent for the original dataset is indicated by a vertical solid red line.

the fundamental differences between the evolutionary regimes of the functional classes of genes that result in the dramatic differences in scaling exponents?

C. The theoretical framework

We seek to uncover the evolutionary roots of the differential scaling of the functional classes of genes within the framework of the general theory of genome evolution by gene gain and loss [4,23–25]. The model used here [Eq. (3)] is identical to the genome evolution model developed previously [18]. However, in the previous studies [18,19], the model was utilized only to account for the evolution of the genome size (total number of genes). Here, we further develop the model and analyze quantitatively the evolution of the genome functional content, i.e., scaling of different functional classes of genes with the genome size, within the same modeling framework. The resulting scaling is given by (see Sec. II C 4 for derivation)

$$x = (1/p_1)x_1 e^{-\Delta S_1(x_1)}, \tag{2}$$

where ΔS_1 and p_1 are the class-specific selection coefficient and genome plasticity, respectively. Both quantities

TABLE II. Model variables and parameters.

Quantity	Description	Expressed by	Equations
x	Number of genes		(3)
x_1	Number of class-specific genes		(5)
P^+	Complete genome gain rate	α and S_0	(3), (8)
P^-	Complete genome loss rate	β and S_0	(3), (9)
P_1^+	Class-specific gain rate	p_1, α, and S_1	(5), (11)
P_1^-	Class-specific loss rate	x, x_1, β, and S_1	(5), (10)
F	Fixation probability	S_0	(7)
α	Acquisition rate		(8)
β	Deletion rate		(9)
S_0	Complete genome selection coefficient	x	(22)
S_1	Class-specific selection coefficient	S_0, ΔS_1	(13)
I	Complete genome pairwise intersection		(14)
k	Complete genome pairwise intersection decay constant	x and P^-	(15)
I_1	Class-specific pairwise intersection		(16)
k_1	Class-specific pairwise intersection decay constant	x_1 and P_1^-	(17)
ΔS_1	Class-specific selection coefficient	q, ξ_1, and x_1	(18)
p_1	Genome plasticity	x_1, a, and b	(11), (19)
q	ΔS_1 slope		(18)
ξ_1	ΔS_1 offset		(18)
a	Genome plasticity intercept		(19)
b	Genome plasticity slope		(19)

(ΔS_1 and p_1) are approximated by the first-order expansion and are extracted from the genomic data as detailed in Sec. II D.

In addition, the modeling framework is extended to account for the divergence of gene repertoires in evolving prokaryotes (Sec. II C 5). All model parameters, including their relations and relevant equations numbers are summarized in Table II, and all modeling assumptions are listed in Table III.

1. Modeling prokaryotic genome evolution

The simplest model for genome size dynamics describes genome evolution as a succession of stochastic gain and loss events [18]. The dynamics of the total number of genes in the genome x is therefore determined by the per-genome gain and loss rates (P^+ and P^-), respectively,

$$dx/dt = P^+ - P^-. \tag{3}$$

In the general case, the gain and loss rates P^+ and P^- depend on the genome size x. A steady-state distribution is formed around the equilibrium genome size [18],

$$P^+ = P^-. \tag{4}$$

Throughout the analysis hereafter, it is assumed that the genome size is approximately constant; that is, the genomes evolve under a long-term equilibrium with respect to gene gain and loss such that Eq. (4) holds. The equilibrium approximation is widely accepted for prokaryotic genome evolution modeling [20,26–29].

To account for the dynamics of distinct functional classes of genes, we define class-specific gain and loss rates. Like the complete genome, each functional class (x_1) is subject to stochastic gains and losses of genes that occur with rates P_1^+ and P_1^-, respectively,

$$dx_1/dt = P_1^+ - P_1^-, \tag{5}$$

TABLE III. Summary of model assumptions.

	Assumption	Model quantities	Equations
1	Genome evolution is dominated by gene loss and HGT	dx/dt, dx_1/dt	(3), (5)
2	Mutations appear and get fixed sequentially	P^+, P^-, P_1^+, P_1^-	(8)–(11)
3	Genome size is in equilibrium	x, x_1, I, I_1	(2), (4), (6), (14), (16)
4	Infinite gene pool $L \gg x$	I, k, I_1, k_1	(14)–(17)
5	Class-specific selection landscape and genome plasticity are similar across all genomes	P_1^+, P_1^-	(10), (11)

with an equilibrium value x_1 that satisfies

$$P_1^+ = P_1^-. \tag{6}$$

In the next subsection, we express gain and loss rates explicitly and show how class-specific rates are related to the overall genome gain and loss rates.

2. Explicit formulation for gene gain and loss rates for finite population

Assuming a finite effective population size under the weak genome dynamics limit (acquisition and deletion rates are low enough such that acquisitions and deletions occur and get fixed sequentially), the gene gain and loss rates can be expressed as the product of the mutation rate and the probability for the mutation to get fixed in the population F [18]. The fixation probability depends on S_0, the genomic mean of the selection coefficient normalized by effective population size [30] (see Appendix D)

$$F(S_0) = \frac{S_0}{1 - e^{-S_0}}. \tag{7}$$

Following the seminal genome size analyses by Lynch and Conery [31], we further assume that the organisms' fitness can be expressed as a function of the number of genes. This assumption implies symmetry of the selective effects with respect to gain and loss of a single gene: The benefit (or cost) is of equal magnitude for gain and loss events but with opposite signs [18,19]. Formally, if acquisition of a gene is associated with a selection coefficient S_0, deletion of the same gene is associated with a selection coefficient $-S_0$. Denoting acquisition and deletion rates by α and β, respectively, the gain and loss rates are

$$P^+ = \alpha(x)F(S_0), \tag{8}$$

$$P^- = \beta(x)F(-S_0). \tag{9}$$

The S_0 value can be regarded as the mean selective benefit (or cost) associated with the acquisition or loss of a random gene. In principle, the mean value of S_0 could be obtained by measuring the selective effect that is associated with a deletion of one gene at a time and averaged over all gene deletions in all genomes and their respective environments. As we explain above, there is symmetry between gain and loss events with respect to the selective effect. However, a closer examination of the gene acquisition process reveals a more complicated picture that involves two distinct timescales. Even genetic material that is beneficial on a large timescale appears to be measurably deleterious initially so that fitness is recovered only after a transient time period of several hundred generations [32]. In contrast, the coefficient S_0 is inferred from extant genomes and thus reflects the average cost (or benefit) of gene deletion, and accordingly, the long-term average benefit (or cost) carried by a gene already incorporated in the genome. Within this framework, the short timescale, that is, the transient phase of gene acquisition, is incorporated into the gain rate of Eq. (8) through the acquisition rate α. Specifically, α represents the combined effect of the DNA insertion rate and HGT barrier, that is, the probability that the acquired gene is not eliminated from the population within the short timescale.

Gain and loss rates for genes that belong to a specific functional class can be expressed following reasoning similar to that used for the complete genome gain and loss rates of Eqs. (8) and (9). The class-specific selection coefficient that determines the fixation probability term can differ from the mean selection coefficient of the complete genome. Under the assumption that deletions occur at random loci across the genome, the class-specific loss rate is given by the complete genome deletion rate β multiplied by the fraction of the genome that is comprised of the genes of a given functional class. Together with the fixation probability of a deletion event that depends on the class-specific mean selection coefficient S_1, this multiplication gives

$$P_1^- = \frac{x_1}{x}\beta(x)F(-S_1). \tag{10}$$

The acquisition rate for class-specific genes is given by the product of the global acquisition rate α, fixation probability that depends on the class-specific mean selection coefficient S_1, and the class-specific genome plasticity p_1,

$$P_1^+ = p_1\alpha(x)F(S_1). \tag{11}$$

Here, p_1 is a modifier for the genomewide acquisition rate that determines the rate for the given gene class. As in the complete genome case, this formulation implies symmetry between class-specific gain and loss, with respect to the selective effect: The selective benefit (or cost) is of equal magnitude for both events but with opposite signs. Accordingly, S_1 quantifies the long-term benefit or cost. If the short-term behavior is similar across all genes, the probability of a successful uptake of a gene is taken into account in the class-specific gain rate by α [Eq. (11)]. In this case, the class-specific acquisition rate is given by the product of α and the fraction of class-specific genes in the external gene pool, so that p_1 simply reflects the class-specific availability of genes. However, as we describe in detail below, an analysis of the scaling laws, together with the pairwise intersection of the gene sets, shows that p_1 is genome-size dependent and does not fit the assumption of a uniform HGT barrier across all classes of genes. The coefficient p_1 therefore quantifies not only the availability of class-specific genes but also the class-specific ability of the microbial cell to tolerate additional genes of the given functional class within the short timescale. Hence, we denote p_1 as class-specific genome plasticity.

3. Selection-drift balance in the evolution of genome size

The relation between the selection coefficient S_0 and the deletion bias β/α under the assumption of a steady state can be obtained by substituting the explicit expressions for P^+ and P^- of Eqs. (8) and (9) into Eq. (4)

$$e^{S_0} = \beta(x)/\alpha(x). \tag{12}$$

Equation (12) quantifies the selection-drift balance with respect to the evolution of the genome size. When gene gain is beneficial and is associated with a positive selection coefficient S_0, equilibrium is possible only when gene deletion is more frequent than acquisition ($\beta > \alpha$), such that the selective pressure towards genome growth is balanced by the intrinsic deletion bias. Similarly, equilibrium in the case when genome growth is counterselected, i.e., $S_0 < 0$, is only possible when acquisitions occur more frequently than deletions ($\beta < \alpha$). Finally, in the special case when acquisition and deletion rates are equal ($\beta = \alpha$), equilibrium is possible only in the strictly neutral case ($S_0 = 0$). As demonstrated by our previous analysis, on average, $S_0 > 0$, which requires a deletion bias to reach equilibrium in genome evolution [18,19]. Indeed, intrinsic deletion bias had been consistently detected for diverse genomes [33–35]. As we show in the following subsection, the relation of Eq. (12) is also useful to relate x and x_1, that is, to obtain the model formulation for the scaling laws.

4. Model prediction for scaling of class-specific genes with genome size

The relation between the number of class-specific genes x_1 and the genome size x [Fig. 5(a)] of Eq. (2) can be obtained by substituting the explicit expressions for P_1^+ and P_1^- of Eqs. (10) and (11) into Eq. (6), together with the relation for S_0 and β/α of Eq. (12). The coefficient ΔS_1 in Eq. (2) is the mean selective (dis)advantage of a gene in the given functional class with respect to a random gene

$$\Delta S_1 = S_1 - S_0. \tag{13}$$

The scaling depends on two factors, class-specific genome plasticity p_1 and class-specific selection coefficient ΔS_1, and can be interpreted as follows. If p_1 is constant, the scaling is determined by ΔS_1 [Fig. 5(b)]. For a constant (that is, independent of the number of genes in the class) ΔS_1, the scaling is linear. Sublinear or superlinear scaling emerges when ΔS_1 depends on the number of genes $\Delta S_1 = \Delta S_1(x_1)$. Specifically, the scaling is sublinear when ΔS_1 decreases with x_1 and superlinear when ΔS_1 increases with x_1 [Fig. 5(c)].

5. Model of genome content evolution

One of the key observable measures of microbial genome evolution is the pairwise intersection between genomes (I), that is, the number of orthologous genes shared by a pair of genomes. Importantly, in this study we extend the modeling framework to account for the evolution of the genome content, and not only the number of genes. As we show below, the model analysis demonstrates that the pairwise intersection decays exponentially with the evolutionary distances, which is incorporated into the modeling framework. Accounting for the evolution of the genome content is a crucial extension of the model with respect to previous studies [18,19] that allows inference of the class-specific selection coefficient from the genomic data, as we explain in detail in the next section.

Both the number of genes in a genome and the pairwise intersections between gene complements result from the same evolutionary processes of stochastic gene gain and loss events. A complete theoretical description of genome evolution should therefore account for both of these quantities. The stochastic gain and loss of genes entail a decay in pairwise genomes similarity through the course of evolution, even when the total number of genes remains approximately constant. As a first-order approximation, given an infinite external gene pool [26], the pairwise genome intersections decay exponentially with the tree distance d (see Appendix E for derivation)

$$I(d) = xe^{-kd}. \tag{14}$$

The rate of pairwise genome similarity decay is determined solely by the gene loss rate, with the decay constant k proportional to the per-gene loss rate

$$k = t_0(P^-/x), \tag{15}$$

where t_0 is a conversion constant from tree distance units to time units. This model fits comparative genomic observations on the pairwise genome similarity decay with evolutionary distance in archaea, bacteria, and bacteriophages [20,36,37]. We test these observations on the ATGC set analyzed in the present work and confirm the close agreement of the model with the data [Fig. 6(a)]. Extraction of the decay constants from the genomic dataset is depicted in Figs. 3(d)–3(f), and the fitting scheme is detailed in Appendix B.

With respect to the genome content, all quantities can be defined for genomic subsets that include only genes from a specific functional class. Similar to its complete genome analog, the class-specific pairwise intersection I_1 (i.e., the number of genes of class 1 shared between the pair of genomes) decays exponentially with evolutionary distance [Figs. 6(b) and 6(c)]

$$I_1(d) = x_1e^{-k_1d}, \tag{16}$$

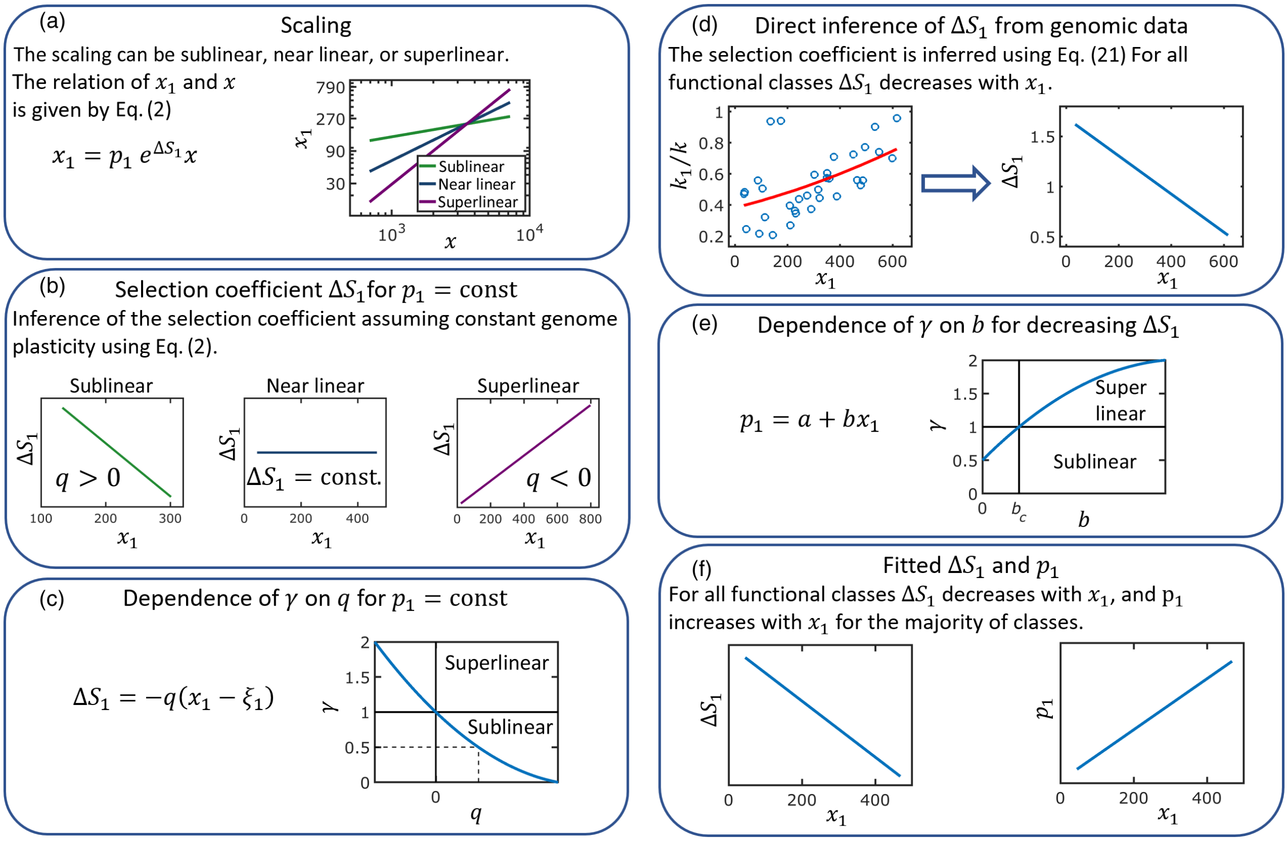


FIG. 5. Visual representation of model scaling according to Eq. (2). The figure presents an outline of the model analysis demonstrating the emergence of genome plasticity. The class-specific selection coefficient ΔS_1, which is inferred from the genomic data, implies class-specific genome plasticity that depends on the number of genes in the given class $p_1 = p_1(x_1)$. (a) An illustration of the scaling of the number of genes in functional classes with the total genome size. Three representative cases are illustrated: sublinear, near-linear, and superlinear scaling. (b) For constant genome plasticity, the scaling is determined by the selection coefficient. Three selection coefficients corresponding to the three scaling exponents in (a) are shown. The sign of q [see Eq. (18)] is indicated for each case. (c) Schematic illustration of the dependence of the scaling exponent γ on q, as implied by Eq. (2). For the constant selection coefficient, the scaling is linear ($\gamma = 1$). The value of q that corresponds to (e) is indicated by a dashed line. (d) For all functional classes, q is positive such that the selection coefficient decreases with x_1. (e) Schematic illustration of the dependence of the scaling exponent γ on the plasticity slope b [see Eq. (19)] under a positive q value [see (c)]. For $b = 0$, the scaling is sublinear, with a γ value identical to the value indicated by a dashed line in (c). For large enough b values, the scaling turns from sublinear to superlinear (indicated by black lines). (f) Illustration of the inferred class-specific selection coefficients and genome plasticity. For all functional classes, the selection coefficient decreases with the genome size, whereas genome plasticity increases with the genome size for the majority of the classes.

where the decay constant k_1 is proportional to the class-specific per-gene loss rate

$$k_1 = t_0(P_1^-/x_1). \tag{17}$$

Empirically, gene classes with sublinear exponents are characterized by slow decay of pairwise intergenome similarity, whereas those with superlinear exponents show fast decay [Fig. 6(d)].

D. Extraction of model parameters from genomic data

After establishing the modeling framework, the next step is to infer the two factors of Eq. (2) that dictate the relation of x_1 and x, that is, the class-specific selection coefficient ΔS_1 and the genome plasticity p_1. Because the scaling laws are robust with respect to local effects and are (nearly) universal across all prokaryotes (see Fig. 1), the evolutionary forces underlying scaling ΔS_1 and p_1 are likely to be universal as well. In particular, nonlinear scaling ($\gamma \neq 1$) suggests that at least one of these factors depends on the number of genes [see Eq. (2)], and we aim to extract from the genomic data a first-order approximation of these dependences. However, the extraction of model parameters involves two major challenges that require construction of a subtle fitting scheme.

First, the selection coefficient determines the class-specific loss rate, whereas the genome plasticity is the principal determinant of the class-specific gain rate. Because the genome size is affected by both gene gain and gene loss, considering merely the number of genes, or the scaling laws for that matter, makes it impossible to infer p_1 without making any assumption on ΔS_1 and vice versa.

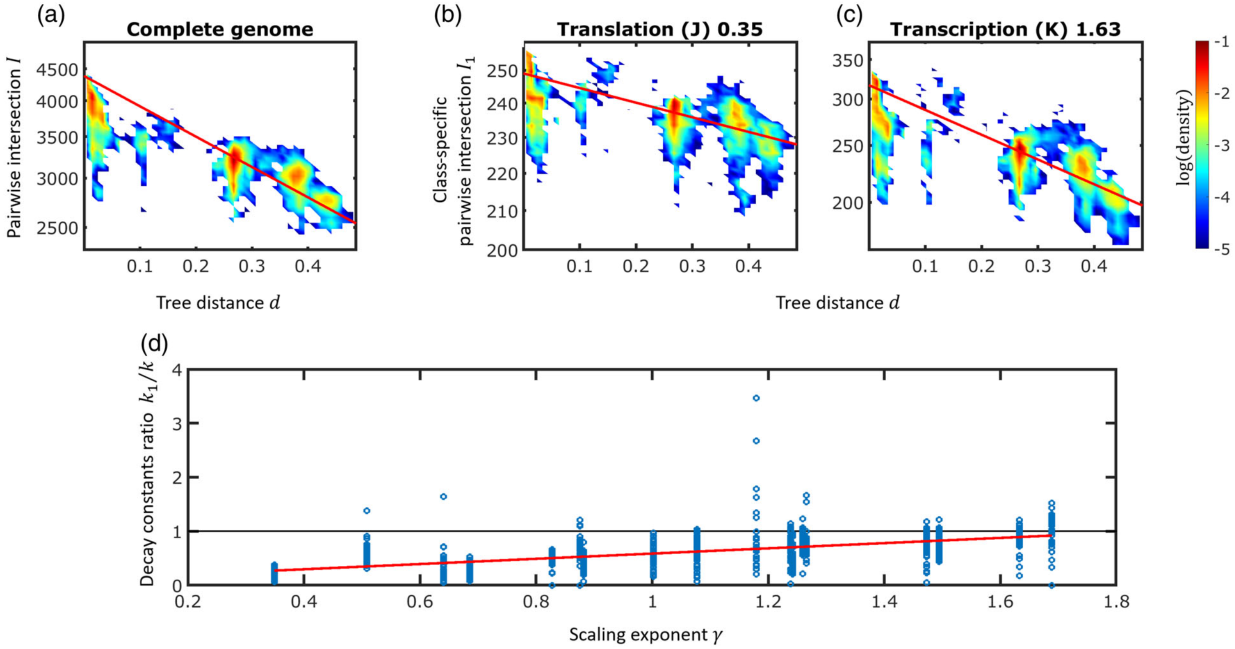


FIG. 6. Decay of prokaryotic gene content similarity with tree distance. The figure illustrates the agreement between the model prediction of the exponential decay of the pairwise intersections between genomes with the evolutionary distance with the genomic data. (a) The pairwise intersection between genomes I plotted against the tree distance d for complete genomes of *E. coli*. Intersections are calculated across all pairs of genomes in the ATGC, and the colors represent the point density where each pair of genomes is represented by a point ($n = 93,096$). The exponential decay fit of Eq. (14) is shown by the red solid line. (b) Pairwise intersections between genomes for translation genes (J) from the *E. coli* genomes. Intersections are calculated for all pairs of genomes in the ATGC, and the colors represent the point density where each pair of genomes is represented by a point. The exponential decay fit of Eq. (16) is shown by the red solid line. (c) Same as (b) for transcription genes (K). (d) Scatter plot showing the class-specific decay constant normalized by the genomewide decay constant vs the scaling exponents of the different functional classes. A linear fit is shown by the red line. Defense genes (V) and the mobilome (X) deviate from the typical values and are excluded from the plot.

We therefore incorporate into the analysis the pairwise genome intersections that are independent of genome plasticity [see Eqs. (15) and (17)] and allow inference of the class-specific selection coefficient. Thus, the genome intersection is a crucial ingredient in the analysis that allows us to disentangle the class-specific selection coefficient and class-specific genome plasticity.

Second, to extract the dependence of either ΔS_1 or p_1 on genome size, it is essential to compare different taxa of different genome sizes. Recently, we have shown that genome evolution is subject to local effects and is governed by taxon-specific factors [19] in addition to the universal factors. To circumvent this taxon specificity represented here by the genomewide acquisition and deletion rates α and β, we normalize the class-specific quantities by the genomic mean quantities for each ATGC separately. This normalization cancels out the ATGC-specific factors and allows us to infer the universal evolutionary factors.

To minimize the number of assumptions and parameters in the model, $\Delta S_1(x_1)$ and $p_1(x_1)$ are approximated by linear functions that can be regarded as first-order expansions of the actual functions

$$\Delta S_1(x_1) = -q(x_1 - \xi_1), \tag{18}$$

$$p_1(x_1) = a + bx_1. \tag{19}$$

Our main objective is to infer from the genomic data the four parameters (q, ξ_1, a, and b) for each functional category. All the inference scheme stages starting from the genomic data to the extraction of model parameters of Eqs. (18) and (19) are shown graphically in Fig. 3.

1. Functional class-specific selection coefficients

The class-specific selection coefficient ΔS_1 is inferred from the ratio between the class-specific decay constant k_1 and the genomic mean k. This ratio can be expressed by class-specific and complete genome loss rates using Eqs. (15) and (17)

$$k_1/k = (x/x_1)(P_1^-/P^-). \tag{20}$$

By substituting the explicit expressions for the loss rates of Eqs. (9) and (10) into Eq. (20), we obtain the relation between the k_1/k ratio and ΔS_1,

$$k_1/k = F[-(\Delta S_1 + S_0)]/F(-S_0). \tag{21}$$

Equation (21) is used to infer from the genomic data the class-specific selection coefficient ΔS_1. The inference stages are depicted in Figs. 3(g)–3(i), and technical details of the optimization procedure are given in Appendix F. The genomewide selection coefficient S_0 is determined based on our previous results [19], where we found that the complete genome selection coefficient S_0 is related to the total number of genes x by

$$S_0 = \ln(0.7x^{0.06}). \qquad (22)$$

Given that we consider the ratio k_1/k, the taxon-specific deletion rate β and the conversion constant t_0 cancel out, such that the ratio depends only on global factors, allowing an unbiased comparison among the ATGCs. The interpretation of the relation between the ratio k_1/k and the selection coefficients above is that genes that are associated with larger selection coefficients are exchanged less frequently than those that are subject to a weaker selection. For example, amino acid metabolism genes (E) show a k_1/k ratio that increases with the number of genes [Fig. 7(a)], suggesting that the fitness cost of deletion of genes in this class drops for larger genomes. This behavior is typical and common to most functional classes, with the notable exception of defense genes (V) and the mobilome (X; the entirety of integrated mobile genetic elements [38]) that show k_1/k greater than 1 for all genome sizes, implying $\Delta S_1 < 0$ for all ATGCs. Accordingly, inferred ΔS_1 decreases with the class-specific number of genes x_1 [Fig. 7(b)]. However, as explained above and illustrated in Figs. 5(b) and 5(c), constant plasticity combined with ΔS_1 that decreases with the genome size results in a sublinear scaling [see Eq. (2)]. The only way to reconcile the decreasing selection coefficient and superlinear scaling is to introduce genome-size-dependent class-specific genome plasticity $p_1 = p_1(x_1)$, as illustrated in Fig. 5(e).

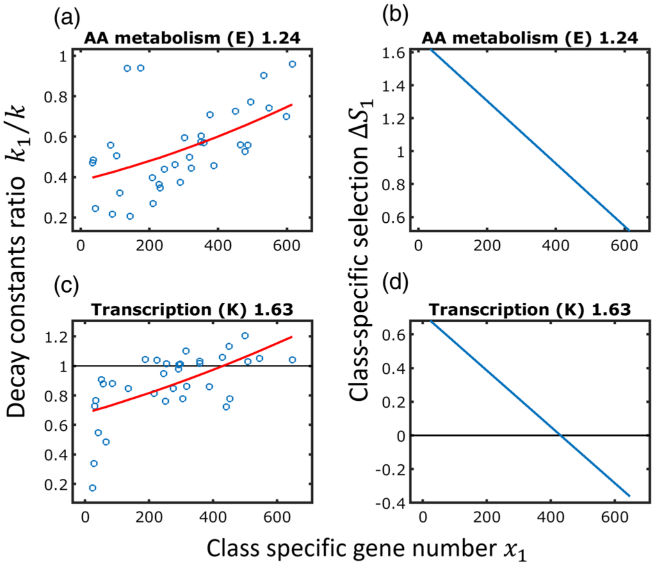


FIG. 7. Inferred selection coefficients for two functional classes of prokaryotic genes. The figure shows the comparison of two functional classes, demonstrating how different pairwise genome similarity decay rates translate into different class-specific selection coefficients. (a) Decay constant ratio k_1/k is plotted against the number of genes in the functional class x_1 for amino acid metabolism genes (E). Each point corresponds to an ATGC from the dataset. The model fit is shown by the solid red line [see Fig. 3(h) for details]. (b) Inferred ΔS_1 for amino acid metabolism genes (E) resulting from the fit shown in (a). (c) Same as (a) but for transcription genes (K). (d) Same as (b) but for transcription genes (K).

2. Functional class-specific genome plasticity

Similar to the inference of ΔS_1 above, to express p_1 by measurable quantities, we use the ratio of k_1 and k. Relying on the equilibrium assumptions of Eqs. (4) and (6), we substitute the loss rates in Eq. (20) for gain rates. Next, we substitute the explicit expressions for gain rates of Eqs. (8) and (11) to obtain

$$(k_1 x_1)/(kx) = p_1 F[(\Delta S_1 + S_0)]/F(S_0). \qquad (23)$$

Similar to Eq. (21), which is used to infer ΔS_1, local influences [19] cancel out. Equation (23) is used to infer the class-specific genome plasticity p_1. The inference stages are depicted in Figs. 3(j)–3(l), and the technical details of the optimization procedure are given in Appendix F.

E. Inferred model parameters

To better understand how the number of genes in each class is determined by the selection coefficient and genome plasticity, it is useful to compare different classes in some detail. For example, for amino acid metabolism genes (E), the k_1/k ratio is below unity [Fig. 7(a)], and accordingly, the fitted ΔS_1 is positive even for larger genomes [Fig. 7(b)]. For this gene class, the inferred plasticity increases with the genome size, leading to the observed moderate superlinear scaling, despite the decrease in ΔS_1 with x_1 [see Fig. 5(e)]. In contrast, the relative abundance of transcription genes (K), primarily, regulators, grows with the genome size such that the k_1/k ratio becomes greater than unity [Fig. 7(c)], which correspond to the fitted ΔS_1 turning negative [Fig. 7(d)]. The higher abundance and the superlinear scaling of transcription genes (K) is therefore attributed to the genome plasticity of this class, which is twice as high as that for amino acid metabolism genes (E) (see Table I). This trade-off between the selection coefficient and genome plasticity is common to all gene classes, and consequently, there is a strong negative correlation between the mean values of inferred ΔS_1 and genome plasticity [Fig. 8; Spearman correlation coefficient $\rho = -0.79$; $p_{\text{val}} < 10^{-3}$ for all functional classes; Spearman correlation coefficient $\rho = -0.77$; $p_{\text{val}} < 10^{-3}$ when omitting the mobilome (X) that demonstrates a significantly higher plasticity than all other functional classes].

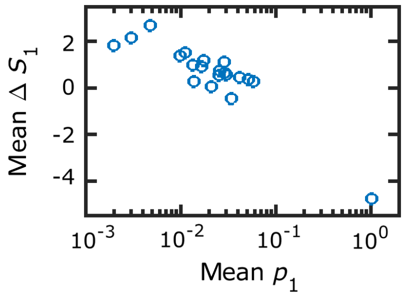


FIG. 8. Correlation between the gene-class-specific selection coefficient and genome plasticity values. The mean inferred ΔS_1 is plotted against the mean inferred plasticity p_1 for all functional classes. The mean values are calculated by averaging over all ATGCs. For all functional classes, ΔS_1 decreases with x_1. However, whereas genes of low-plasticity classes are difficult to acquire and losses of these genes are rare and incur a large selective cost, genes of high-plasticity classes are acquired frequently, and conversely, the loss of these genes typically incur only a low cost. Consequently, a strong negative correlation between the mean selection coefficient and genome plasticity is observed across the functional classes of genes [Spearman correlation coefficient $\rho = -0.79$; $p_{\mathrm{val}} < 10^{-3}$ for all functional classes; Spearman correlation coefficient $\rho = -0.77$; $p_{\mathrm{val}} < 10^{-3}$ when omitting the mobilome (X) that demonstrates extreme values of genome plasticity and selection coefficient].

Finally, we test the model consistency by reconstructing the scaling laws using the fitted selection coefficients and genome plasticity. Specifically, for each gene class, the fitted selection coefficient and genome plasticity are substituted into Eq. (2) (Fig. 1). For most classes, the fit quality of our model is comparable to, albeit slightly worse than, that of the power-law fit (Table S1 in the Supplemental Material [39]). The immediate sources of errors in model fitting are the linear approximations for ΔS_1 and p_1 of Eqs. (18) and (19). It should be noted that the fitted scaling is obtained from a population-genetics model rather than as a fit of an arbitrary function. Moreover, model parameters are inferred not only from the number of genes but from the combination of the genomewide or class-specific number of genes and the pairwise gene content similarity decay rates in ATGCs [Eqs. (21) and (23); see Figs. 3(h)–3(l)] which carry complementary information and allow one to extract the selection coefficients. Both the numbers of genes in a complete genome and in each functional class and the pairwise similarity decay rates are readily measurable quantities that characterize genome evolution. For all functional classes, with the exception of the defense systems (V) and the mobilome (X), the relative selection coefficient is positive and decreases with the genome size [Fig. 5(f), Table I]. For all except three functional classes (L, replication and repair; D, cell division; V, defense), genome plasticity increases with the number of genes [Fig. 5(f), Table I]; that is, the larger the genome, the higher the probability that an additional gene can be incorporated into the corresponding functional networks.

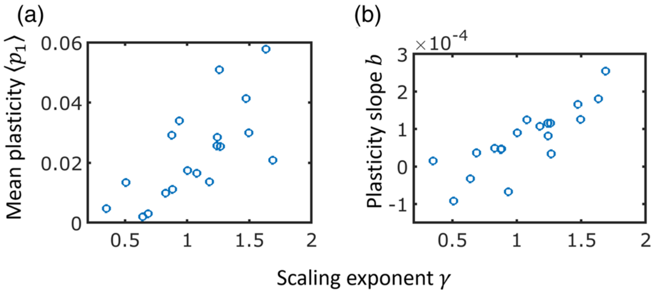


FIG. 9. Correlation between scaling exponents and genome plasticity values for different functional classes of genes. The scaling exponent strongly correlates with the genome plasticity, implying that, to a large extent, genome plasticity dominates prokaryotic genome evolution. (a) Mean plasticity across all ATGCs is plotted against the scaling exponent. Each point corresponds to a functional class of genes. The mobilome is associated with genome plasticity that is an order of magnitude greater than those of the other gene classes and is excluded from the plot. (b) The plasticity slope is plotted against the scaling exponent. Similar to (a), each point corresponds to a functional class of genes, and the mobilome is excluded from the plot.

Both the plasticity slope and the mean plasticity strongly positively correlate with the scaling exponent with the respective Spearman correlation coefficients $\rho = 0.81$ $(p_{\mathrm{val}} < 10^{-3})$ and $\rho = 0.74$ $(p_{\mathrm{val}} < 10^{-3})$ (Fig. 9). This strong correlation suggests that genome plasticity, together with the selection coefficient, shape the evolution of genome content.

III. DISCUSSION

In this work, we develop a general theoretical model explaining the universal scaling of the functional classes of genes in prokaryotes (Fig. 2). The scaling that we obtain from this simple model does not follow a power law exactly but gives a comparable quality of fit within the range of available data, even if slightly inferior to direct power-law fits. However, it should be stressed that the model parameters' inference does not rely only on the number of genes but incorporate additional information derived from pairwise similarity in gene content, which is a crucial ingredient that allows inference of the selection coefficient, without any assumptions on genome plasticity. The model does not include any assumptions on specific relationships between different functional classes as postulated in the previous models [10]. Instead, we introduce an additional class-specific parameter, which we denote genome plasticity, that is distinct from the selection coefficient and, together with the latter evolutionary factor, governs gene gain and loss processes. Genome plasticity reflects the availability of the genes of the given functional class, which itself depends on their abundance in the external gene pool, as well as the strength of purifying selection, on the short timescale, against horizontally acquired genes that has been previously described as the HGT barrier [40]. The difference from the selection coefficient is that the selective component of genome plasticity corresponds to short-term selection, whereas the selection coefficient applies to substantially longer timescales. Optimization of the model parameters in our previous study indicated that the dependence of the deletion bias on the genome size is weak, $\beta/\alpha \propto x^{0.06}$ [19]. This finding implies that the dependence of the gene acquisition rate on genome size is similar to that of the deletion rate on genome size, which is often taken as linear [26,28].

Our current results provide a biologically plausible explanation for the dependence of the gene acquisition rate on the genome size, namely, that the genome size increase enhances genomic plasticity, such that the HGT barrier is lowered. Our analysis shows that genes from different functional classes are acquired at widely different rates, with distinct dependences on genome size. Accordingly, inferred genome plasticity differs across different functional classes, which correlates with the scaling exponents (Fig. 9) and implies class-specific HGT barriers. This finding relates previous observations that genes of different functional classes undergo HGT at different rates [41–43] to the scaling laws.

The biological explanation of the differences in plasticity among functional classes of genes, at least, in part, could lie in the so-called complexity hypothesis [44]. This hypothesis postulates that genes encoding components of complex, interconnected functional systems are less likely to be transferred horizontally than genes coding for proteins that function in comparative isolation. Indeed it has been shown that HGT rates show negative correlation with the number of protein interactions [41,45]. It is therefore plausible that genome plasticity is strongly affected by the connectivity of the genes in each functional class. More generally, plasticity can be considered one of the forms of evolvability, a much debated concept [46–50] that, however, becomes the key factor shaping genome evolution in our model. It should be emphasized that genome plasticity, as introduced in the present model, endows evolvability with a precise mathematical form. Functional classes of genes with high plasticity, and accordingly, superlinear scaling exponents, are evolutionarily flexible and can be thought of as the microbial adaptation resource. The biological features of these classes appear compatible with this interpretation. Indeed, the four gene classes with the highest scaling exponents, namely, secondary metabolism (Q), transcription (K), signal transduction (T), and carbohydrate metabolism (G), are involved in the response of bacteria to rapidly changing environmental cues, including various biological conflicts (many of genes in the Q class are involved in antibiotic production and resistance). These classes have high (G and K) or moderate (Q and T) plasticity and accordingly can accumulate in genomes to the point that the class-specific relative selection coefficient ΔS_1 becomes negative so that these genes incur a non-negligible fitness cost on the organism. The genome similarity decay constant ratio k_1/k for these functional classes is unity or greater in the majority of the ATGCs; that is, these genes are also lost at rates similar to or higher than the average gene, resulting in their overall dynamic evolution. Notably, the gene classes with only a general functional prediction (R) and without any prediction (S) also show superlinear scaling (albeit less pronounced than the above four classes) and high plasticity, suggesting that at least some of these genes contribute to adaptive processes. In agreement with previous results [51], we find that defense systems and the mobilome incur a fitness cost on prokaryotes, and the relative cost of the mobile elements is an order of magnitude greater than that of defense systems. Not surprisingly, the genome plasticity of the mobilome also stands out, being at least an order of magnitude greater than those of all other classes (Table I). Conversely, for sublinear classes, plasticity is low, so that incorporation of additional genes is unlikely, albeit becoming more accessible in larger genomes. The genes in these classes are responsible for housekeeping functions that contribute less to short-term adaptation than the superlinear gene classes.

Several simplifying assumptions are made throughout the derivation to allow the theoretical analysis and to keep the model tractable (Table III). In particular, the class-specific selection coefficient represents the average over all genes of the given class in all genomes and over long evolutionary spans. This is an obvious simplification, and indeed, although for 15 of the 20 functional classes of genes, the optimal set of parameters is found to be highly robust; for the remaining five classes, optimization fails to converge on a single set of optimal parameters (Fig. S1 in the Supplemental Material [39]). This instability might reflect the functional heterogeneity of these classes of genes (which is apparent, for example, in the case of the mobilome) as well as complex patterns of gene gain and loss which could reflect changes of the selection coefficient with time and/or environment fluctuations. Regarding the gene acquisition process, it is assumed that genes are gained one at a time and from an infinite gene pool (that is, no repeated gene gain). Furthermore, it is assumed that genomes are in stochastic equilibrium in terms of the genome size. These simplifying assumptions notwithstanding, nonequilibrium reacquisitions of the same genes or gain of more than one gene at a time cannot explain the different acquisition rates that are observed for different functional classes. Under our model, the differences in gene gain rates are determined by the class-specific genome plasticity, a key parameter of genome evolution that has not been explicitly introduced previously.

As a characteristic of the evolution of gene classes that can be directly determined from genome comparison and does not depend on any model of evolution, we analyze the class-specific core genomes and pangenomes [52–54] (Fig. 10). The normalized core genome size for the

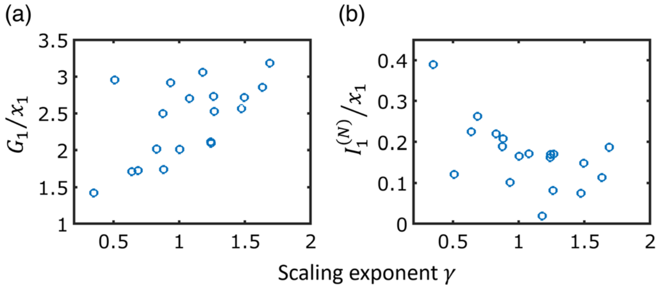


FIG. 10. Scaling exponents for different functional classes of genes, core genomes, and pangenomes. The class-specific pangenome G_1 (a) and core genome $I_1^{(N)}$ (b) are plotted against the scaling exponent for *E. coli*. Each point corresponds to a functional class of genes. To allow comparison between classes, pangenomes and core genomes are normalized by the number of genes in each class. The plots show that the scaling exponent strongly and positively correlates with the class-specific pangenome size but strongly and negatively correlates with the class-specific core genome size. Thus, for the sublinear classes, such as translation, the difference between the pangenome and the core genome is minimal, whereas for the superlinear classes, such as transcription, the pangenome dramatically exceeds the core genome, reflecting high plasticity.

individual functional classes of genes correlates with the scaling exponent significantly and negatively (Spearman correlation coefficient coefficients $\rho = -0.55$; $p_{\text{val}} = 0.007$). In contrast, the normalized class-specific pangenome size correlates with the scaling exponent significantly and positively, with Spearman correlation coefficient coefficients $\rho = 0.56$ ($p_{\text{val}} = 0.005$). Thus, as expected, the sublinear classes have large relative core genomes and small relative pangenomes, in contrast with the superlinear classes that make the principal contribution to the pangenome expansion. The results presented here indicate that class-specific genome plasticity is the principal determinant of gene gain and, accordingly, the evolutionary factor that shapes the dynamics and architecture of microbial pangenomes and the process of rapid adaptation in microbes.

ACKNOWLEDGMENTS

We thank Koonin group members for helpful discussions. This work utilized the computational resources of the NIH HPC Biowulf cluster [55], and was supported by the NIH Intramural Research Program at the National Library of Medicine.

APPENDIX A: GENOMIC DATASET

Clusters of closely related species from the ATGC database [21] that contain ten or more genomes each are used in the analyses. The database includes fully annotated genomes and a phylogenetic tree for each cluster. The phylogenetic trees' branch length units are substitutions per site, and trees are inferred from a concatenation of nucleotide sequence alignment of core genes in each ATGC (see Ref. [21] for more details). Within each cluster of genomes, genes are grouped into clusters of orthologs (ATGC COGs). Out of all genome clusters that contain ten genomes or more, we select the 36 genome clusters that match the following criteria: (i) maximum pairwise tree distance is at least 0.1, and (ii) the ATGC is not composed of two groups of tightly related genomes, such that pairwise tree distances are centered around more than two typical values (see Fig. S2 in the Supplemental Material [39]). Two of the 36 genome clusters are identified as outliers and are excluded from the dataset (see Fig. S2 and Table S2 in the Supplemental Material [39]). The 34 genome clusters analyzed in this study are listed in Table S3 in the Supplemental Material [39]. The ATGC COGs are assigned to functional categories as defined in the COG database [22]. It should be stressed that COGs and ATGC COGs represent two different extreme cases: Whereas the COG dataset is constructed for a broad prokaryotic diversity (large phyletic depth), the ATGC COGs in each ATGC are constructed for closely related genomes (small phyletic depth). The genome sizes and sizes of the functional classes of genes are given by the number of ATGC COGs that are present in each genome and belong to the respective classes. Multiple genes from a single genome that belong to the same ATGC COG are counted once. The behavior of the genes without orthologs in other genomes (ORFans) deviates from the genomic mean, in particular, due to their extremely high turnover rate [20]. Therefore, ORFans are excluded from the present analyses. A genome content analysis is performed for 20 COG categories. Functional classes of genes that are analyzed are listed in Table I.

APPENDIX B: DATA FITTING AND OPTIMIZATION OF MODEL PARAMETERS

The numbers of genes in each class are discrete counts that typically span about 1 order of magnitude. Because the data span a large range and there is no justification to assume homoscedasticity of errors, the fitting cannot be performed by optimizing the coefficient of determination, which assumes that errors follow a normal distribution. It is therefore assumed that the errors follow a negative binomial distribution, which accounts for different error dispersions. Specifically, fitting is performed by optimizing model parameters together with the negative binomial distribution dispersion parameter, such that the log-likelihood is maximal.

1. Inference of scaling exponent

Power-law scaling exponents are obtained by fitting the genomic data to the power law of Eq. (1). For each functional class, parameters η and γ together with the negative binomial distribution dispersion parameter are optimized by maximizing the log-likelihood for all genomes in the dataset. Genomes that do not contain genes that belong to the respective class are excluded from the analysis. The resulting fits are shown in Fig. 1, and the fit Akaike information criterion values are listed in Table S1 of the Supplemental Material [39].

2. Inference of pairwise intersection decay constants

The pairwise intersections decay constants k and k_1 are inferred by fitting Eqs. (14) and (16) separately for each ATGC to the genomic data. Since ORFans are omitted from the dataset, the intercept is set to the mean number of genes (x for complete genomes and x_1 for class-specific genes), such that the decay constant and the negative binomial dispersion parameter are optimized by maximizing the log-likelihood. Genomes that do not contain genes that belong to the respective class are excluded from the analysis.

APPENDIX C: STATISTICAL ANALYSIS OF SCALING EXPONENTS

For each functional class, a power law is fitted to a collection of genes generated by bootstrapping the original dataset. Specifically, the sampled dataset is generated by sampling with replacement the ATGCs and collecting all genomes in sampled ATGCs. Sampling is performed over ATGCs and not directly at the level of genomes in order to avoid sampling bias due to the different number of genomes in each ATGC. The distribution of the fitted scaling exponents is shown for each class for 1000 bootstrap samplings in Fig. 4. For each pair of classes, the distribution overlap C is calculated and shown in Table S4 of the Supplemental Material [39]. Specifically, for categories X and Y, with scaling exponents $\gamma^X \leq \gamma^Y$ for the original dataset and bootstrap exponents γ_i^X and γ_j^Y, the overlap is given by

$$C_{XY} = \left(\sum_{i=1}^{1000}\sum_{j=1}^{1000} c_{ij}^{XY}\right)/1000^2 \tag{C1}$$

with

$$c_{ij}^{XY} = \begin{cases} 1 & \text{for } \gamma_i^X > \gamma_j^Y, \\ 0 & \text{else.} \end{cases} \tag{C2}$$

Given that, for the original dataset, the scaling exponent of class X is smaller than that of class Y, the overlap C_{XY} indicates the probability of a bootstrap exponent of class X to be greater than the bootstrap exponent of class Y. Accordingly, $C_{XX} = 1/2$.

APPENDIX D: POPULATION-SIZE-NORMALIZED SELECTION COEFFICIENT

Scaling the time by the effective population size N_e allows us to express gain and loss rates through $S_0 = N_e s_0$, where s_0 is the genomewide average of the selection coefficient. Substituting into the genome size dynamics of Eq. (3) the gain and loss rates of Eqs. (8) and (9), we get the explicit form

$$dx/dt = \alpha(x)F(S_0) - \beta(x)F(-S_0), \tag{D1}$$

where F denotes the fixation probability of Eq. (7).

APPENDIX E: PAIRWISE GENOME INTERSECTIONS

To account for the genome content similarity, each genome is represented by a vector $\boldsymbol{X}$ with elements that assume values of 1 or 0. Each entry represents an ATGC COG, where 1 or 0 indicate the presence or absence, respectively, of that ATGC COG in the genome. Genome size x is then given by the sum of all elements in $\boldsymbol{X}$. The number of common genes I is defined as

$$I(t) = \langle \boldsymbol{X} \cdot \boldsymbol{Y} \rangle, \tag{E1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two vectors that represent the two genomes, the angled brackets indicate averaging over all possible pairs of genomes, and the dot operation stands for a scalar product. The pairwise genomes intersection dynamic is given by

$$dI/dt = 2\langle (d\boldsymbol{X}/dt) \cdot \boldsymbol{Y} \rangle, \tag{E2}$$

where we use the fact that both averages are equal $\langle (d\boldsymbol{X}/dt) \cdot \boldsymbol{Y} \rangle = \langle \boldsymbol{X} \cdot (d\boldsymbol{Y}/dt) \rangle$. For a finite gene pool of size L (the limit of an infinite gene pool, which is used in the Results section, is calculated below), we have

$$\langle (d\boldsymbol{X}/dt) \cdot \boldsymbol{Y} \rangle = -P^- \frac{L}{L-x} I(t)/x + P^- \frac{x}{L-x}, \tag{E3}$$

where the last approximation relies on the steady-state assumption $P^+ \approx P^-$. Substituting the relation above into the equation for the pairwise genome similarity time derivative of Eq. (E2) and solving the differential equation, we obtain the exponential decay of the pairwise genome intersection to an asymptote x^2/L,

$$I(t) = [I(0) - x^2/L]e^{-\nu t} + x^2/L \tag{E4}$$

with decay constant

$$\nu = \frac{2P^-}{x}\frac{L}{L-x}. \tag{E5}$$

The asymptote x^2/L can be obtained using a simple intuitive derivation. The two intersecting genomes are regarded as two independent random samples of x genes from a pool of L genes. The probability that a sample from the second genome includes a gene that was already sampled in the first genome is x/L. To obtain the number of intersecting genes, this probability is multiplied by the number of trials, that is, the number of genes x, giving the intersection asymptote x^2/L. Assuming a clock with respect to loss events, the time t can be translated into the tree pairwise distance as $d = 2t/t_0$. Further assuming that the gene pool is much larger than the mean genome size $L \gg x$ (formally equivalent to the assumption of an infinite gene pool [26]), we get for I the exponential decay of Eq. (14), with the decay constant of Eq. (15).

Finally, it is possible to consider pairwise genome intersections with respect to a subset of genes. The derivation presented in Eqs. (E1)–(E5) can be repeated for genes that belong to a specific functional class, resulting in an exponential decay of the class-specific pairwise intersections I_1 of Eq. (16), with the decay constant k_1 of Eq. (17).

APPENDIX F: OPTIMIZATION OF MODEL PARAMETERS

For each functional class, four model parameters q, ξ_1, a, and b of Eqs. (18) and (19) are optimized using the mean numbers of genes and decay constants for each ATGC x, x_1, k, and k_1. Specifically, all four model parameters are optimized simultaneously using Eqs. (20) and (22), together with S_0 of Eq. (21), by maximizing the goodness of fit R^2 for both equations. The model parameters are optimized by maximizing a goal function that is given by the sum of goodness-of-fit values for both equations. In principle, four-dimensional optimization can converge at a local minimum or else it could have substantially different nearly optimal solutions. To ensure that optimal parameters are picked and to demonstrate the robustness of the optimal solution, the following procedure is applied. In the first stage, optimization is performed for each functional class starting from an arbitrary point in the parameter space

$$\begin{aligned} a &= \mathrm{mean}(x_1/x),\\ b &= 0,\\ q &= 0.01,\\ \xi_1 &= \max(x_1). \end{aligned} \tag{F1}$$

The outcome of this optimization, including both the optimized parameters and the goal function values, is set as a benchmark, where each functional class is associated with different values. Further optimizations are performed starting from random points in the parameter space, and we keep the results of 100 optimizations with goal function values equal to or greater than the benchmark value G_0. For each functional class, the differences in the parameters' values in each of the 100 optimizations and the benchmark optimization are calculated. The difference D is calculated for each model parameter as

$$D = \left|\frac{z_0 - z_i}{z_0}\right|, \tag{F2}$$

where z_0 denotes the benchmark value of a model parameter, and z_i is an optimal model parameter value obtained when starting the optimization from a random point in the parameter space. For all but five functional classes (cell division D, defense V, secretion N, mobilome X, and function unknown S), optimizations converge consistently, with negligible D, as shown in Fig. S1 of the Supplemental Material [39]. The large differences observed for five classes imply that there are two or more local maxima with goal function values equal to or larger than G_0.

In the next stage, we find the global maximum. A procedure similar to the one described above is applied, but instead of setting the benchmark based on an arbitrary starting point, the benchmark goal function value is taken as the maximum value of 100 optimizations that start from random points in the parameter space. Optimizations are then performed starting from random points in the parameter space until 100 solutions with the goal function value equal to or greater than the benchmark value are obtained. In this case, each of the 20 functional classes converge to a single-class-specific point in the parameter space, as shown in Fig. S3 of the Supplemental Material [39]. All five categories (D, V, N, X, and S) that previously showed large differences between the optimal and near-optimal parameter values converge to a solution with a negative q [positive slope of the class-specific selection coefficient; see Eq. (18)]. For these five classes, we apply the constraint $q > 0$ and repeat the search. The differences for the optimal and near-optimal solutions with $q > 0$ are shown in Fig. S4 of the Supplemental Material [39]. The genome plasticity can be determined, but different starting points converge at different selection coefficients. The solution with the largest goal function value is taken as the optimal solution, and comparison of the selection coefficients and genome plasticities for optimal solutions with $q > 0$ and $q < 0$ are shown in Fig. S5 of the Supplemental Material [39].

[1] C. K. Stover *et al.*, *Complete Genome Sequence of Pseudomonas aeruginosa PAO1, an Opportunistic Pathogen*, Nature (London) **406**, 959 (2000).

[2] E. van Nimwegen, *Scaling Laws in the Functional Content of Genomes*, Trends Genet. **19**, 479 (2003).

[3] K. T. Konstantinidis and J. M. Tiedje, *Trends between Gene Content and Genome Size in Prokaryotic Species with*

Larger Genomes, Proc. Natl. Acad. Sci. U.S.A. **101**, 3160 (2004).

[4] E. V. Koonin and Y. I. Wolf, *Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World*, Nucleic Acids Res. **36**, 6688 (2008).

[5] N. Molina and E. van Nimwegen, *Scaling Laws in Functional Genome Content across Prokaryotic Clades and Lifestyles*, Trends Genet. **25**, 243 (2009).

[6] E. De Lazzari, J. Grilli, S. Maslov, and M. C. Lagomarsino, *Family-Specific Scaling Laws in Bacterial Genomes*, Nucleic Acids Res. **45**, 7615 (2017).

[7] O. X. Cordero and P. Hogeweg, *Regulome Size in Prokaryotes: Universality and Lineage-Specific Variations*, Trends Genet. **25**, 285 (2009).

[8] S. Maslov, S. Krishna, T. Y. Pang, and K. Sneppen, *Toolbox Model of Evolution of Prokaryotic Metabolic Networks and Their Regulation*, Proc. Natl. Acad. Sci. U.S.A. **106**, 9743 (2009).

[9] T. Y. Pang and S. Maslov, *A Toolbox Model of Evolution of Metabolic Pathways on Networks of Arbitrary Topology*, PLoS Comput. Biol. **7**, e1001137 (2011).

[10] J. Grilli, B. Bassetti, S. Maslov, and M. C. Lagomarsino, *Joint Scaling Laws in Functional and Evolutionary Categories in Prokaryotic Genomes*, Nucleic Acids Res. **40**, 530 (2012).

[11] B. O. Bengtsson, *Modelling the Evolution of Genomes with Integrated External and Internal Functions*, J. Theor. Biol. **231**, 271 (2004).

[12] E. V. Koonin, K. S. Makarova, and L. Aravind, *Horizontal Gene Transfer in Prokaryotes: Quantification and Classification*, Annu. Rev. Microbiol. **55**, 709 (2001).

[13] C. Pal, B. Papp, and M. J. Lercher, *Adaptive Evolution of Bacterial Metabolic Networks by Horizontal Gene Transfer*, Nat. Genet. **37**, 1372 (2005).

[14] T. J. Treangen and E. P. Rocha, *Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes*, PLoS Genet. **7**, e1001284 (2011).

[15] P. Puigbo, A. E. Lobkovsky, D. M. Kristensen, Y. I. Wolf, and E. V. Koonin, *Genomes in Turmoil: Quantification of Genome Dynamics in Prokaryote Supergenomes*, BMC Biol. **12**, 66 (2014).

[16] W. F. Doolittle, *Lateral Genomics*, Trends Cell Biol. **9**, M5 (1999).

[17] N. Molina and E. van Nimwegen, *The Evolution of Domain-Content in Bacterial Genomes*, Biol. Direct **3**, 51 (2008).

[18] I. Sela, Y. I. Wolf, and E. V. Koonin, *Theory of Prokaryotic Genome Evolution*, Proc. Natl. Acad. Sci. U.S.A. **113**, 11399 (2016).

[19] I. Sela, Y. I. Wolf, and E. V. Koonin, *Estimation of Universal and Taxon-Specific Parameters of Prokaryotic Genome Evolution*, PLoS One **13**, e0195571 (2018).

[20] Y. I. Wolf, K. S. Makarova, A. E. Lobkovsky, and E. V. Koonin, *Two Fundamentally Different Classes of Microbial Genes*, Nat. Microbiol. **2**, 16208 (2016).

[21] D. M. Kristensen, Y. I. Wolf, and E. V. Koonin, *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 (2017).

[22] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin, *Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Satabase*, Nucleic Acids Res. **43**, D261 (2015).

[23] A. R. Mushegian and E. V. Koonin, *A Minimal Gene Set for Cellular Life Derived by Comparison of Complete Bacterial Genomes*, Proc. Natl. Acad. Sci. U.S.A. **93**, 10268 (1996).

[24] A. B. Kolsto, *Dynamic Bacterial Genome Organization*, Mol. Microbiol. **24**, 241 (1997).

[25] E. V. Koonin, *Comparative Genomics, Minimal Gene-Sets and the Last Universal Common Ancestor*, Nat. Rev. Microbiol. **1**, 127 (2003).

[26] F. Baumdicker, W. R. Hess, and P. Pfaffelhuber, *The Infinitely Many Genes Model for the Distributed Genome of Bacteria*, Genome Biol. Evol. **4**, 443 (2012).

[27] D. H. Huson and M. Steel, *Phylogenetic Trees Based on Gene Content*, Bioinformatics **20**, 2044 (2004).

[28] R. E. Collins and P. G. Higgs, *Testing the Infinitely Many Genes Model for the Evolution of the Bacterial Core Genome and Pangenome*, Mol. Biol. Evol. **29**, 3413 (2012).

[29] P. Marttinen and W. P. Hanage, *Speciation Trajectories in Recombining Bacterial Species*, PLoS Comput. Biol. **13**, e1005640 (2017).

[30] D. M. McCandlish, C. L. Epstein, and J. B. Plotkin, *Formal Properties of the Probability of Fixation: Identities, Inequalities and Approximations*, Theor. Popul. Biol. **99**, 98 (2015).

[31] M. Lynch and J. S. Conery, *The Origins of Genome Complexity*, Science **302**, 1401 (2003).

[32] S. Bershtein, A. W. Serohijos, S. Bhattacharyya, M. Manhart, J. M. Choi, W. Mu, J. Zhou, and E. I. Shakhnovich, *Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria*, PLoS Genet. **11**, e1005612 (2015).

[33] D. A. Petrov, T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw, *Evidence for DNA Loss as a Determinant of Genome Size*, Science **287**, 1060 (2000).

[34] D. A. Petrov, *DNA Loss and Evolution of Genome Size in Drosophila*, Genetica **115**, 81 (2002).

[35] C. H. Kuo and H. Ochman, *Deletional Bias across the Three Domains of Life*, Genome Biol. Evol. **1**, 145 (2009).

[36] G. Plata, C. S. Henry, and D. Vitkup, *Long-Term Phenotypic Evolution of Bacteria*, Nature (London) **517**, 369 (2015).

[37] T. N. Mavrich and G. F. Hatfull, *Bacteriophage Evolution Differs by Host, Lifestyle and Genome*, Nat. Microbiol. **2**, 17112 (2017).

[38] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint, *Mobile Genetic Elements: The Agents of Open Source Evolution*, Nat. Rev. Microbiol. **3**, 722 (2005).

[39] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.9.031018 for Supplemental figures and tables.

[40] R. Sorek, Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork, and E. M. Rubin, *Genome-Wide Experimental Determination of Barriers to Horizontal Gene Transfer*, Science **318**, 1449 (2007).

[41] O. Cohen, U. Gophna, and T. Pupko, *The Complexity Hypothesis Revisited: Connectivity Rather Than Function Constitutes a Barrier to Horizontal Gene Transfer*, Mol. Biol. Evol. **28**, 1481 (2011).

[42] Y. Nakamura, T. Itoh, H. Matsuda, and T. Gojobori, *Biased Biological Functions of Horizontally Transferred Genes in Prokaryotic Genomes*, Nat. Genet. **36**, 760 (2004).

[43] R. Merkl, *A Comparative Categorization of Protein Function Encoded in Bacterial or Archeal Genomic Islands*, J. Mol. Evol. **62**, 1 (2006).

[44] R. Jain, M. C. Rivera, and J. A. Lake, *Horizontal Gene Transfer among Genomes: The Complexity Hypothesis*, Proc. Natl. Acad. Sci. U.S.A. **96**, 3801 (1999).

[45] A. Wellner, M. N. Lurie, and U. Gophna, *Complexity, Connectivity, and Duplicability as Barriers to Lateral Gene Transfer*, Genome Biol. **8**, R156 (2007).

[46] A. Wagner, *Robustness, Evolvability, and Neutrality*, FEBS Lett. **579**, 1772 (2005).

[47] A. Crombach and P. Hogeweg, *Evolution of Evolvability in Gene Regulatory Networks*, PLoS Comput. Biol. **4**, e1000112 (2008).

[48] M. Pigliucci, *Is Evolvability Evolvable?*, Nat. Rev. Genetics **9**, 75 (2008).

[49] J. Masel and M. V. Trotter, *Robustness and Evolvability*, Trends Genet. **26**, 406 (2010).

[50] J. Lehman and K. O. Stanley, *Evolvability is Inevitable: Increasing Evolvability without the Pressure to Adapt*, PLoS One **8**, e62186 (2013).

[51] J. Iranzo, J. A. Cuesta, S. Manrubia, M. I. Katsnelson, and E. V. Koonin, *Disentangling the Effects of Selection and Loss Bias on Gene Dynamics*, Proc. Natl. Acad. Sci. U.S.A. **114**, E5616 (2017).

[52] D. Medini, C. Donati, H. Tettelin, V. Masignani, and R. Rappuoli, *The Microbial Pan-Genome*, Curr. Opin. Genet. Dev. **15**, 589 (2005).

[53] G. Vernikos, D. Medini, D. R. Riley, and H. Tettelin, *Ten Years of Pan-Genome Analyses*, Curr. Opin. Microbiol. **23**, 148 (2015).

[54] J. O. McInerney, A. McNally, and M. J. O'Connell, *Why Prokaryotes Have Pangenomes*, Nat. Microbiol. **2**, 17040 (2017).

[55] http://hpc.nih.gov.