Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers

Jacopo Grilli[1], Mariacristina Romano[2], Federico Bassetti[3] and Marco Cosentino Lagomarsino[4,5,*]

[1]Dipartimento di Fisica e Astronomia "G. Galilei", Università di Padova, Via Marzolo 8, I-35131 Padova, Italy, [2]Dipartimento di Fisica, Università degli Studi di Milano, via Celoria, 16, 20133 Milano, Italy, [3]Università di Pavia, Dipartimento di Matematica, via Ferrata 1, 27100 Pavia, Italy, [4]CNRS, UMR 7238, Paris, France and [5]Sorbonne Universités, UPMC Université Paris 06, UMR 7238 Computational and Quantitative Biology, Genomic Physics Group, 15 rue de l'École de Médecine, Paris, France

ABSTRACT

Prokaryotes vary their protein repertoire mainly through horizontal transfer and gene loss. To elucidate the links between these processes and the cross-species gene-family statistics, we perform a large-scale data analysis of the cross-species variability of gene-family abundance (the number of members of the family found on a given genome). We find that abundance fluctuations are related to the rate of horizontal transfers. This is rationalized by a minimal theoretical model, which predicts this link. The families that are not captured by the model show abundance profiles that are markedly peaked around a mean value, possibly because of specific abundance selection. Based on these results, we define an abundance variability index that captures a family's evolutionary behavior (and thus some of its relevant functional properties) purely based on its cross-species abundance fluctuations. Analysis and model, combined, show a quantitative link between cross-species family abundance statistics and horizontal transfer dynamics, which can be used to analyze genome 'flux'. Groups of families with different values of the abundance variability index correspond to genome sub-parts having different plasticity in terms of the level of horizontal exchange allowed by natural selection.

INTRODUCTION

While the traditional view of protein evolution focused solely on amino-acid changes, it is now clear that at a higher level of complexity, the proteome evolves using protein domains as elementary structural and functional modules. These building blocks are combined and arranged over evolutionary times, giving rise to gene families (1–3). In bacteria, the modular expansion and contraction of protein families generates highly variable gene repertoires (4,5). Expansion of families is associated with the acquisition of novel functions and novel regulatory structures. Focusing on the protein/domain families found on the same genome, it is well known that they follow a universal distribution with a cutoff correlated with the genome size (6), which is one among other simple but remarkable statistical regularities (7,8). This particular law is essentially due to the interplay of family-expansion and innovation moves (9–11).

On the other hand, the cross-genomic family abundance statistics is comparatively poorly characterized, and is the subject of our attention here. Families expand predominantly by horizontal gene transfer (HGT) or internal gene duplication. HGT may add radically new genetic information, and liberate genomic evolutionary processes from tinkering exclusively with pre-existing genes (4,5,12). Horizontal transfer is believed to be the dominant component of genome innovation for bacteria (13). Cross-, and intra-species HGT are believed to occur at very high rates, leading to very large pan-genomes (14). Conversely, duplicated genes are fixed by processes of sub- or neofunctionalization, which render both copies adaptive. Ample evidence shows that duplication has an important role in bacterial adaptation, for example in genome adaptive expansion (15). Detected duplications in the *E. coli* clade appear to be recent, and preferentially associated to highly expressed genes (13).

Followed across genomes, the gene content variability is often summarized as a 'gene frequency distribution', which describes, within a set of sequenced genomes, how many single-copy orthologous genes are found to occur in a given fraction of the repertoired genomes (16–20). The gene frequency distribution contains relevant information about the evolutionary mechanisms that shape the gene content of

*To whom correspondence should be addressed. Tel: +33(0)1 44277341; Fax: +33 (0)1 4427336; Email: marco.cosentino-lagomarsino@upmc.fr

a species' genome, but focuses solely on average occurrence, rather than on the variability across genomes. Here, we approach the statistics for cross-species gene-content variation in terms of gene-family abundance.

In this case, besides the presence/absence of a family, one has to deal with the inter-species variability of the number of members of that family (which we term here 'abundance') across sequenced genomes. In other words, each family is characterized by a specific 'fluctuation' pattern of its abundance across genome rather than by a single mean observable as in the case of gene frequency. These cross-genomic gene-family histograms, or 'gene-family abundance distributions', are shaped by events of gene duplication, loss and horizontal transfer.

Using data concerning domain (super)families for 1065 sequenced distinct bacterial species, we find a link between the amount of horizontal transfers within a family and the width of its abundance histogram. We use a minimal stochastic model to rationalize this result. The model predicts an increased dispersion in family abundance as intra-species family expansion (comprising duplication and intra-species horizontal transfer) becomes more relevant over inter-species horizontal transfer. Additionally, we find that a third class of gene families exists, characterized by a family abundance histogram that is more sharply peaked than possibly expected by the model. Analyzing the functional categories of the families that follow this trend, we argue how they could be subject to selection relative to gene abundance. These findings allow us to define an index able to classify some evolutionary properties of gene families purely based on their cross-species abundance histograms, and different genome sub-parts that behave differently in terms of the levels of horizontal exchange and duplication allowed by natural selection.

MATERIALS AND METHODS

A collisional model for cross-species gene-family evolutionary dynamics

In the model, a fixed number N of species' genomes interact pairwise randomly, and can gain genes by 'collisions' associated with HGT events; the collisions are assumed to be homogeneous, but we will see that this assumption can be relaxed with no effect on the predictions. At each time step, two species are chosen randomly and they can exchange genes by HGT by drawing them from each other's genome with Bernoulli trials of probability p_h. During the same time, they draw in a similar way from their own genome a set of genes to be lost (with probability p_l) and duplicated (with probability p_d). The mathematical formalization of this process is described in the Supplementary Data. Hence, a time step ideally represents the characteristic time between two fixed inter-species horizontal transfer events. Draws from different families are assumed to be independent (see the main text and Supplementary text). To obtain stationarity it is technically necessary to assume that $p_h + p_d = p_l$ (in addition to $p_d + p_l \leq 1$), and thus the total number of elements $\sum_{i=1}^{N} V_i$ of all families across genomes is, on average, conserved (see Supplementary Text). The model was accessed

through both direct simulations and mean-field analytical calculations.

Data sources

The domain compositions of the proteins of all analyzed bacterial genomes and the superfamily functional annotations were retrieved from the SUPERFAMILY database (21) (release 1.75 27/1/2013), which contains protein domain superfamily assignments using the SCOP (22) structural classification of proteins, for all protein sequences in completed genomes. We considered the 1065 bacterial genome species in the data set, and excluded the strains, as their presence would bias the abundance profiles. We also analyzed separately a set of 225 unicellular eukaryotes merged with data from 62 higher eukaryotes for which the domain assignment of the longest transcript per gene (excluding splicing variants) were available from the SUPERFAMILY database. For the analysis of functional annotations, we considered the function annotation of SCOP domain superfamilies, which is a scheme of 50 detailed functional categories, mapped to seven more general function categories, developed by C. Vogel (23).

Horizontal transfers inferred by the following three independent methods were considered. (i) The Horizontal Gene Transfer Database (HGT-DB (24)), which scores genes that differ statistically from the rest of the genome in G+C content and/or amino-acid usage. (ii) The Dark-Horse database (25), which scores phylogenetically atypical genes on a genome assigning a 'lineage probability index' (LPI); we considered the set genes having scores higher than the threshold LPI > 0.6 and excluded matches with reciprocal LPI scores <0.75. (iii) The data set from (13), which detected, using phylogenetic methods, gene-family expansion by both horizontal transfer and duplication. For each data set, the analysis was restricted to the set of genes present in the SUPERFAMILY database. The final data set (i) contains a total of 549 genomes and 51 658 domains, data set (ii) a total of 348 genomes and 21 004 domains, and data set (iii) a total of 20 genomes and 1864 domains. Transferred (or duplicated) domains were defined as the assigned SUPERFAMILY domains of a transferred (duplicated) gene. Plasmid family enrichment was computed using the domain assignments of the NCBI plasmid sequence collection available within the SUPERFAMILY database, and the 'unusual families' tool of the database web site.

Data analysis

Genomes were divided into sliding bins based on their size in assigned domains. Each bin contained a size interval of 390 domains; for this size interval, we verified that the mean and the variance of domain superfamily abundance was stable when sliding the bins. The sampling weight $w_{f,b}$ of a (super)family f, for each bin b, is defined considering the number of genomes in the bin n_b and the number of nonzero entries n_b^+ of the family in the bin, $w_{f,b} = n_b n_b^+ / (\sum_b n_b^2)$. Each family is assigned weight $w_f = \sum_b w_{f,b}$. To calibrate the analysis, we performed a preliminary survey of the abundance histograms for metabolic domain superfamilies. We found that of 318 families, 66 showed abundance distributions that

were very similar to Poisson. The rest showed different behaviors. We used this result to define a 'noise threshold', to identify the empirical histograms that are sampled sufficiently well, as the value of the sampling weight w_f for the 66th-ranking family ($w_f \simeq 0.38$). The order parameter L_f is defined as $L_f = \sum_b L_{f,b} n_b^+ / (\sum_b n_b^+)$, where, for each bin, the $L_{f,b}$ is the L1 distance of the family abundance histogram in the bin $p_{f,b}(v)$ with a Poisson distribution Poiss(v), $L_{1,f} = \frac{1}{2} \sum_v |p_{f,b}(v) - \text{Poiss}(v)|$. Sums over v were truncated at $v = 450$. The order parameter Q_f is defined as $Q_f = \sum_b Q_{f,b} w_{f,b}$, where

$$Q_{f,b} = (1 - \delta_{\text{Var}_b(v_f),0}) \log \frac{\langle v_f \rangle_b}{\text{Var}_b(v_f)} +$$
$$+ \quad \delta_{\text{Var}_b(v_f),0} \left(\max_b \log \frac{\langle v_f \rangle_b}{\text{Var}_b(v_f)} \right).$$

In the above expression, $\delta_{k,l}$ is Kronecker's delta, *Var* stands for variance, and $\langle$ stands for the mean. $Q_{f,b}$ evaluates the deviation from Poisson behavior in terms of (log) mean-to-variance ratio (inverse Fano factor), considering separately the case of zero variance, which occurs in empirical data. For the analysis of horizontal transfers, the parameter H_f was defined as a mean fraction of transfers over the family size as follows:

$$H_f = \frac{1}{O_f} \sum_g \frac{H_{f,g}}{V_{f,g}},$$

where the index $g = 1$. O_f runs on genomes where the family f is present, $H_{f,g}$ is the number of scored horizontal transfers for family f in genome g, and $V_{f,g}$ the abundance of family f in genome g.

RESULTS

Model simulations and calculations

We discuss first the stochastic model (Figure 1A), since the results are useful to introduce the data analysis. The model describes a minimal dynamics of duplication/loss and interspecies HGT, and formulates a minimal informed expectation for the family abundance profile. The model only describes events that are visible on the representative genome of the species (because they are fixed), and recapitulates the action of selection in the rates p_d, p_h and p_l. Importantly, when compared to data, the model only describes interspecies events, and thus the 'duplication' move is an intraspecies family expansion that includes duplication as well as intra-species horizontal transfers. For simplicity, we will mainly refer to the move as duplication in the description of the model, and explicitly address the question when dealing with the data. Finally, we assume independence between gene families. Thanks to the latter condition, the gene abundance V_i of a single family across all $i = 1...N$ species can be described separately from the others. Note however that, while matching the model with empirical data, the effective rates are allowed to vary from family to family, giving rise to the observed diversity between families, hence this simplifying assumption is not restrictive. Model time maps to evolutionary time in a complex way. In comparing with data, we will assume that observed species had the time to reach a steady state where the gene-family abundance distributions are roughly invariant (i.e. that the stationary abundance distribution is the empirically relevant quantity). The main observable is the family abundance profile, the distribution of the family population V. Using mean-field kinetic equations similar to Boltzmann equations (26), it is possible to estimate the stationary-state value of all moments of V. Processes of the type considered have already been applied in various interdisciplinary contexts (27–30).

In presence of horizontal transfer and loss the model predicts a stationary Poisson distribution

Setting $p_d = 0$, the only effective parameter is $p_h = p_l$. In this regime, the dynamics of a family is dominated by horizontal transfer and loss. Since in the simulated dynamics HGTs are balanced with losses, the mean abundance λ is fixed by the initial conditions. Assuming that these averages are determined by extrinsic or ancestral processes, we ask which kind of abundance profile can be maintained in the hypothetical scenario of 'steady' evolution where duplications (and intra-species HGT gains) are irrelevant. Simulations show that, starting from an arbitrary initial condition, the system relaxes to a stationary state where the family abundance profiles $P(V = v)$ do not vary with time. The stationary abundance histogram resembles a Poisson distribution with parameter λ (Figure 1B). Importantly, when $p_d = 0$, we prove analytically that the mean-field steady-state solution is the Poisson distribution of parameter λ (see Supplementary Text). Note that this prediction is parameter-free, as the only parameter of the Poisson distribution is fixed by the average (Figure 1B).

When duplication is relevant the model predicts overdispersed stationary distributions

When $p_d > 0$ and duplication (or intra-species transfers) cannot be neglected, all three parameters are relevant to define the steady state. Simulations show that increasing p_d makes the steady-state abundance distribution increasingly dispersed (Figure 1C), as confirmed by analytical mean-field calculation of the moments of the distribution (see Supplementary Text). At steady state, we obtain $\text{Var}(V) = \lambda(1 + p_d/(p_h(1 - p_h)))$, which perfectly agrees with simulations (Figure 1C). A closed analytical expression for the full distribution can be obtained in the so-called 'grazing collisions' limit (31), i.e. when $p_d \to 0$ and $p_h \to 0$ with the same rate, and is a negative binomial distribution of mean λ and variance $\lambda(p_d + p_h)/p_h$ (see Supplementary Text).

The relaxation times of all moments of V can be estimated analytically in the mean-field limit (Supplementary Figure S1). In general, the steady state is approached exponentially with relaxation time $1/(p_h(1 - p_h))$, measured in 'sweeps' of N collisions (see Supplementary Text and Figure S1). This result agrees well with finite-N simulations. As we mentioned, model time has a complex relationship with evolutionary time. However, the relevant aspect is that a certain number of fixed horizontal exchanges are required to build up the steady distribution (the relaxation time depends only on p_h, even when $p_d > 0$).
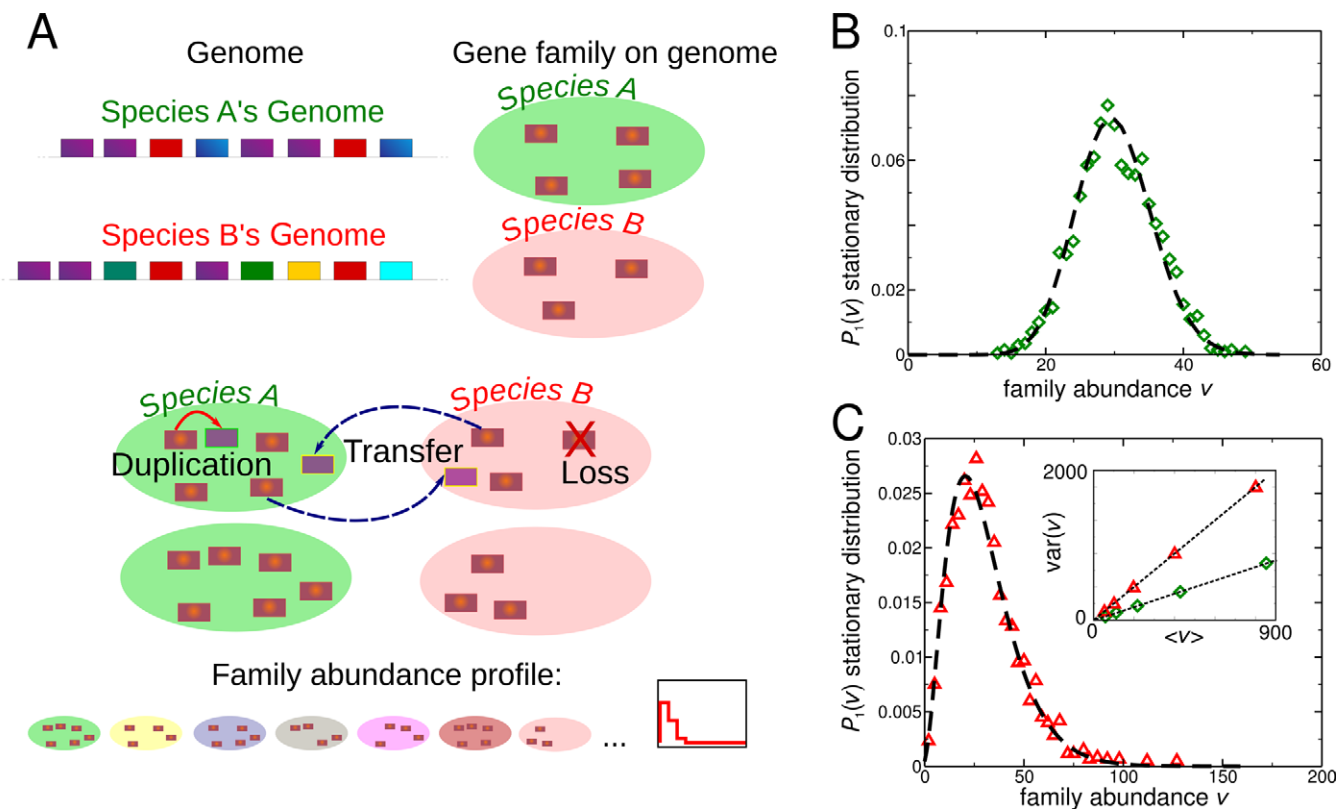
Figure 1. Definition and main predictions of the model. (**A**) Sketch of the model. For each genome representative of a species (top), the model describes the dynamics of each gene family separately over a set of species, represented here (for each species) as the Venn diagram of the elements of the family. Following stochastic binary 'collision' events between species (middle), genes of a given family are exchanged between species, and, over the same time scale, can be duplicated or lost in each species. (**B**) In absence of gene duplications, the model predicts a Poisson distribution for the family abundance profile. Symbols (green diamonds) are the steady-state abundance histogram from simulation of 1000 species, with $p_d = 0$ $p_h = 0.01$ and initial abundance of 30 for all species. The dashed line is the analytical prediction (a Poisson distribution with average 30). (**C**) In presence of duplications, the dispersion of the steady-state abundance histogram increases. Symbols (red triangles) represent the steady-state abundance histogram from simulation of 1000 species, with $p_d = 0.009$ $p_h = 0.001$ and initial abundance of 30 for all species. The dashed line is the analytical prediction, a negative binomial distribution, valid in the limit of small p_d and p_h (see text). The inset shows that in both cases (same symbols as above) the analytical estimates (dashed lines) capture well the scaling of the variance of the abundance profile with the average abundance found in simulations (symbols).

The range of behaviors predicted by the model is found in empirical data

The model contains a number of radical assumptions, but guides the analysis of the empirical gene-family abundance distributions. As we will see, the distributions predicted by the model are present in a relevant fraction of the empirical families and correspond very well to their expected biological and evolutionary features.

We have considered data of domain assignments relative to 1065 sequenced bacteria of different species from the SUPERFAMILY database, which provides a coarse-grained definition of a domain family/superfamily, and thus allows to obtain well-sampled per-family abundances. We performed the main analyses on a total of 1530 superfamilies, and verified that the results were consistent at the family taxonomy level. For each superfamily (or family), indexed by f, we computed the abundance profiles $P^f(v)$. Note that this approach is complementary to the customary approach of phylogenetics studies, where the main focus are presence–absence patterns of orthologs. The abundance counts include orthologs, but mostly xenologs (horizontal transfers) and paralogs (duplicates), and thus in principle contains information about the processes (described by the model) through which new genes are added on a genome. Thus, abundance and presence/absence patterns report about very different information.

It is well-known that family abundance is related to genome size (11), and also that the total population of families with specific biological functions has definite scaling with genome size (32–34). Furthermore, the amount of horizontal transfers increases with overall family size (35). Thus, in order to bypass possible spurious signals, we computed family abundance histograms for sets of genomes belonging to (sliding) bins of similar genome size, measured in domains. This procedure can be justified theoretically, using a model variant that accounts for different genome sizes in a simplified way (see Supplementary Text).

Additionally, many families are sparsely (or poorly) populated, and genome size bins of constant width contain a highly variable amount of genomes. In order to compensate for these sampling problems, we used the weights w_f, which measures the efficiency of sampling.

A survey of the families with well-sampled abundance profiles (high w_f) shows that the model phenomenology is
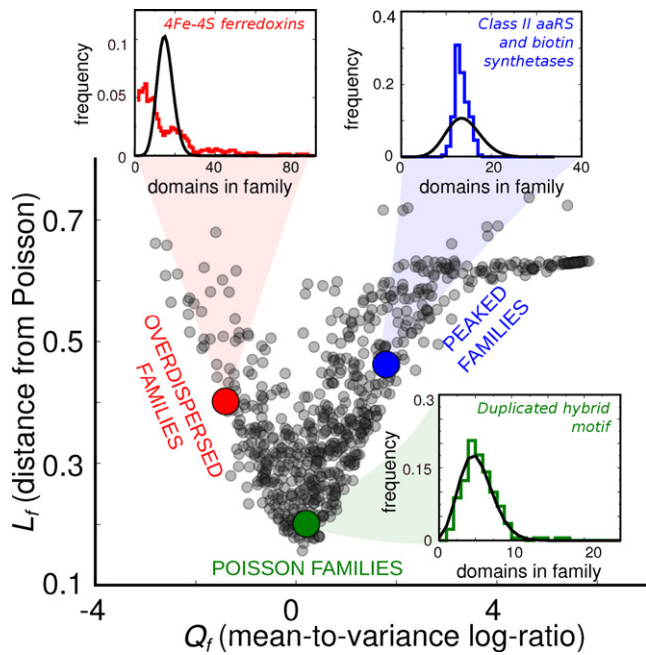
Figure 2. Typical behaviors of the domain (super)family abundance profiles and their classification. Three main behaviors exist for the families, shown by the insets. Each inset compares an example of family abundance histogram (steps) with the Poisson distribution with equal mean (black lines). 'Poisson-like' family profiles correspond to this expectation, while 'overdispersed' and 'peaked' profiles are characterized by larger and lower variance, respectively. The scatter plot represents two different ways to score the deviation from Poisson-like behavior. Each point represents a domain (super)family. The *x*-axis reports the order parameter Q_f, which quantifies the deviation from the equality of mean and variance by their log ratio. The *y*-axis is the parameter L_f, based on the *L*1 distance between the abundance histogram and the Poisson distribution with equal average. Both observables are weighted on the sampling noise.

present in the data (Figure 2; some further representative examples are given in Supplementary Figure S2). Specifically, some families have abundance profile histograms that are very close to Poisson distributions, and some others have a higher dispersion. A third type of family abundance profile is not captured by the model, as it is less dispersed than the Poisson distribution with equal average. A few of these families have zero variance (i.e. they are deterministically populated by the same number of domains across genomes).

Order parameters and behaviors of the gene-family abundance distributions

To further quantify these observations, we have used the statistics, or 'order parameters', L_f and Q_f. According to the model prediction, a family with null behavior, and whose evolution is dominated by horizontal transfers, should have an abundance histogram that looks like a Poisson distribution, and whose average is equal to the mean. The order parameters L_f and Q_f measure the deviation from these two properties (in particular, Q_f is equivalent to an inverse Fano factor, see the Materials and Methods section). Both Q_f and L_f should be close to zero if the abundance profile complies to the model prediction for the case of negligible duplications, a Poisson distribution. The scatter plot of L_f versus

Q_f (Figure 2) shows that this property is valid for a set of families.

The same plot in Figure 2 allows to distinguish families based on their abundance profiles: in the data, abundance profiles deviate from Poisson behavior in two ways: by increasing their variance (i.e. going to negative Q_f) or by decreasing it (positive Q_f). The families that increase it, also increase their dispersion, and are, in principle, captured by the model variant that includes duplication. The families whose variance in the abundance profile is decreased escape the null model, suggesting that they might be subject to further constraints limiting their abundance fluctuations.

The families with lower sampling noise (high w_f) tend to have lower L_f at equal Q_f and to collapse around a common $L_f(Q_f)$ curve (Figure 3A). In particular, for $Q_f \simeq 0$ their abundance profiles resemble more, and more consistently across genome sizes, a Poisson distribution. The collapse of the data with increasing sampling efficiency can be considered an important consistency check with the model, which predicts common underlying abundance profiles. We also found that the sampling weight w_f linearly correlates very strongly with the family occurrence, i.e. the fraction of genomes where family *f* is found (Figure 3A). The occurrence is the domain family equivalent of the so-called 'gene frequency' of orthologs (16–20), and follows a similar distribution. Conversely, the abundance fluctuation index Q_f does not show any clear correlation with the occurrence, indicating that in the data abundance fluctuations and occurrence are distinct (Figure 3B). In addition, the analysis is very robust with respect to domain taxonomy (SCOP families versus superfamilies, Supplementary Figure S3).

Importantly, the fluctuation index Q_f is consistent for groups of genomes with different size (Supplementary Figure S4), while the abundances typically scale with size. In other words, with varying genome size, the index Q_f robustly classifies a scaling of the variance with the mean, which seems to be a family-specific feature. This supports the conclusion that family abundance fluctuations essentially follow three main trends, which are quantified by the index Q_f: 'Poisson-like' abundance profiles tend to follow a Poisson distribution, while 'overdispersed' and 'peaked' profiles have, respectively, increased and decreased species-to-species variability. Notably, the same classification is less consistent for abundance fluctuations in eukaryotic genomes (Supplementary Figure S5), where horizontal transfers are much rarer, and the underlying picture suggested by the model is not expected to readily apply.

The above analysis indicates that, assuming the model, the cases $p_d = 0$ and $p_d > 0$ should be well distinguished by Q_f, and that the distributions are Poisson-like or not exactly according to the model prediction following the values of this parameter. However, we did not attempt any precise parameter inference. The model has three parameters, and the size conservation assumption reduces them to two. However, if $p_d = 0$, the value of p_h cannot be easily inferred because Q_f is insensitive to this parameter. Additionally, in the grazing collision limit, the only recognizable parameter is p_d/p_h. In general, additional knowledge of the skewness of the abundance profiles should in principle allow to estimate p_d, but in the empirical data the histograms appear too noisy to perform this operation reliably.
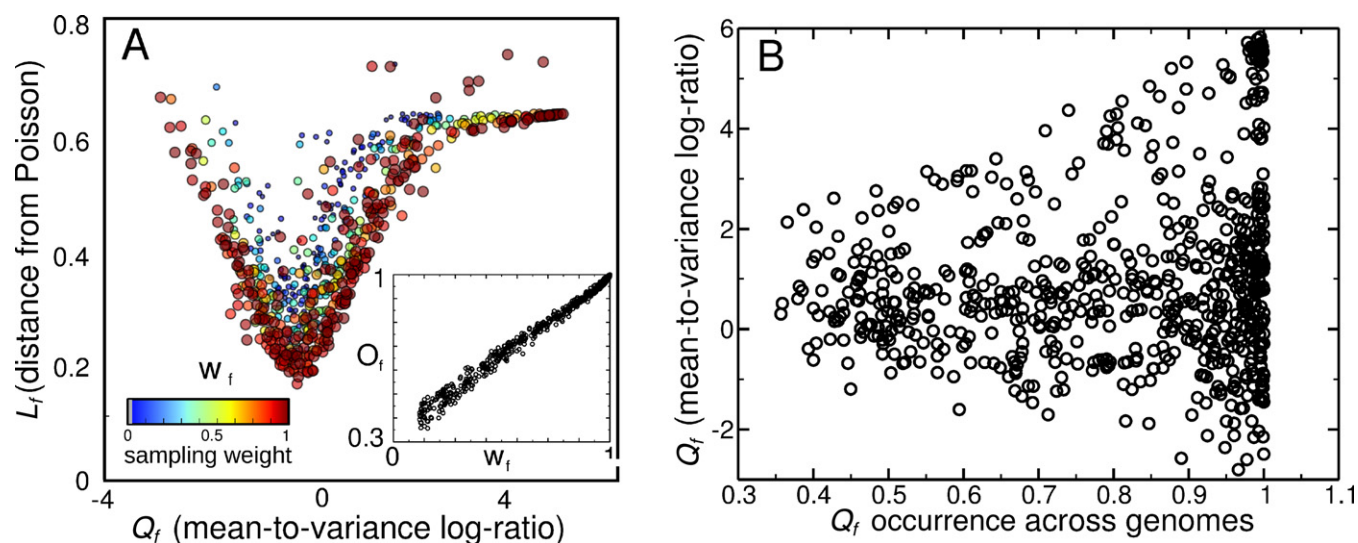
Figure 3. Robustness of the abundance profile classification with respect to the sampling weight of families w_f. (**A**) The scatter plot is the same as the one in Figure 2 of the main text, but color and size of each point correspond to the sampling weight w_f. The 'V' shape of the relation between L_f and Q_f is more marked for the families with high w_f. Inset: linear relation between sampling weight w_f and the occurrence O_f of families. The latter is defined as the fraction of genomes having at least one domain of the family considered. The fact that high w_f families are not preferentially associated with any value of Q_f indicates that the classification operated by Q_f is robust. (**B**) Lack of correlation between O_f and Q_f.

Evolutionary behavior and gene-family abundance profile

The model suggests that different values of Q_f should also correspond to different rates of cross-species HGTs and duplications/intra-species HGTs. To obtain a direct proof that abundance fluctuations are connected with rates of fixed HGTs and gene duplications, we employed four bioinformatics data sets that measured these quantities on a large set of genomes, and compared the amounts of scored transfers and duplication with the index Q_f. We converted the gene-based data into domains using the simple criterion that all domains belonging to transferred or duplicate genes are scored as transferred or duplicate domains.

The first data set, HGT-DB (24), scores putative horizontal transfers on a set of 959 species' genomes using GC content. For each family and genome, this data source estimates the putative genes, and hence domains, gained by transfers, which can be subdivided into families. Each family yields a certain number of transfers per genome. For each family and genome, we normalized the number of horizontally transferred domains to the family size. We defined the parameter H_f of a given family as the average across genomes of this quantity. The cross-genomic average compensates in part the obviously larger sampling errors for families with lower abundance. H_f quantifies the relative contribution of HGT to a family; it approaches 1 when the family has many putative transfers consistently across genomes, and 0 when very few transfers are typically found compared to the family size. Figure 4 shows that families with Poisson-like and overdispersed abundance profiles ($Q_f \lesssim 0$) tend to have higher H_f, and hence carry more horizontal transfers. This is in line with the model predictions (considering that intra-species transfers and duplications are not distinguished by the model). Additionally, we repeated the analysis considering data from the DarkHorse database (25), which scores horizontal transfers on genome-wide basis by a comple-

mentary approach, based on phylogenetic information from the NCBI taxonomy database (36), and the results are very robust (Supplementary Figure S6).

Third, we tested the correlation of Q_f with the enrichment of a family on plasmids as a proxy for the tendency of the family members to be found on 'mobile elements'. To perform this analysis, we used the domain assignments of the NCBI plasmid sequence collection. We found that families with $Q_f \lesssim 0$ tend to be more enriched on plasmids, while families with increasingly larger positive values of Q_f are increasingly underrepresented (Figure 5), in line with the idea that abundance fluctuations are related to family 'mobility' across genomes.

To assess the role of transfer and duplication in gene-family expansion, we analyzed the data set from Treangen and Rocha (13), who estimated the relative contributions of horizontal transfer and duplication to gene-family expansion in a set of closely related complete genomes. In this case, we found results that are consistent with the above conclusions, but the small size of the data set did not allow a thorough statistical analysis (Supplementary Figure S7). In particular, the families that were scored as containing abundant gene duplications consistently have $Q_f < 0$, as expected from the null model.

Finally, in order to further test whether families with different abundance variability build up genome sectors with differential plasticity, we defined three pairwise genome distances based on superfamily usage (37) restricted to families with overdispersed, Poisson-like and peaked abundance profiles, respectively. For each genome pair in the data set, we thus obtained three distances, which we compared to a reference distance obtained from a 16S ribosomal RNA phylogenetic tree (38). While family-based metrics are usually a good signal of phylogeny, HGTs should 'scramble' the signal from vertical descent. In line with this, the genome distances based on Poisson-like and overdispersed fam-
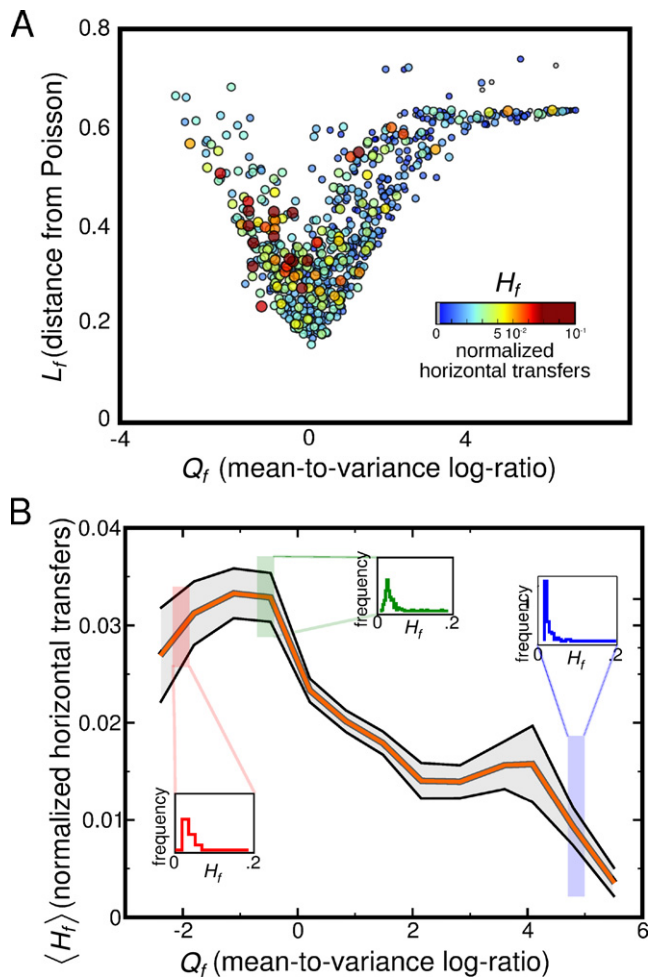
Figure 4. Independent estimates of horizontal transfers within a family correspond to the model expectation from the family abundance profiles. (**A**) The scatter plot is the same as the plot in Figure 2, but the color and size of each point correspond to the parameter H_f, which is an average number of horizontally transferred domains in that family (estimated here using data from the HGT-DB database), normalized, for each genome where family f is found, to the family size. Points with $H_f = 0$ are in gray. Compatibly with the expectation of the model, many horizontal transfers are found for Poisson-like families, i.e. toward the minimum of this plot. For $Q_f > 0$ (peaked) families tend to have high or null H_f. (**B**) Plot of the average of H_f over classes for bins of Q_f. H_f increases with decreasing Q_f, indicating that an increasing number of transfer events are found for families with Poisson-like and overdispersed abundance profiles, compatibly with the null model expectations. The insets show the histograms of H_f for the region connected to them. See Supplementary Figures S6 and S7 for related tests using different data sets.

ily content are less correlated with phylogenetic distance (Spearman ρ 0.36 and 0.32, respectively) compared to the distance based on peaked families (Spearman ρ 0.46).

Biological function and gene-family abundance profile

As well known, the evolutionary properties of a gene family are connected to its functional landscape. Thus, if the classification of gene families based on their abundance histograms is meaningful, one expects a connection between biological function and abundance profiles of the families. We performed Fisher exact tests on the contingency of dif-
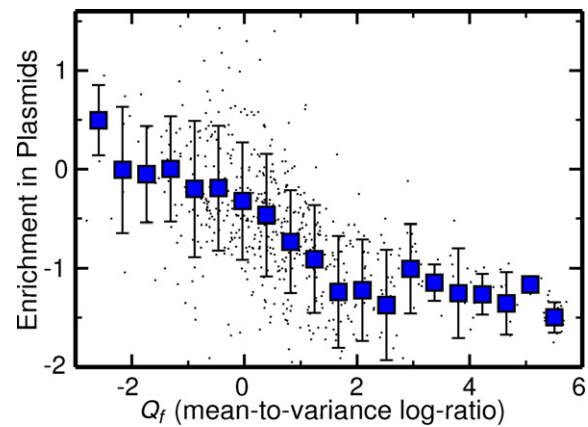
Figure 5. Enrichment of dispersed families on mobile elements (plasmids). The abundances of domain superfamilies on plasmids were obtained from domain assignments of the NCBI plasmid sequence collection. The plot shows the correlation of Q_f with enrichment of families on plasmids. The *y*-axis represents a log ratio score (21) between the percentage of total plasmid domains belonging to a given superfamily and the background percentage computed on all bacterial genomes. This score increases for families that are increasingly enriched on plasmids.

ferent functional categories and abundance profile classes (Table 1). To avoid biases from undersampled families, we restricted the analysis to the subset of 701 (super)families that fall below a 'noise threshold' value of w_f. In order to perform the analysis, we divided these families according to their abundance fluctuations into the following three classes: 'overdispersed' ($Q_f < -0.43$, 130 families), 'peaked' ($Q_f > 1$, 281 families) 'Poisson-like' ($-0.43 < Q_f < 1$, 282 families).

We used the SCOP superfamily domain functional annotations, which comprise 50 detailed functional categories, mapped to seven broader categories: information, regulation, metabolism, intracellular processes, extracellular processes, general, other/unknown. The associations with these broader functional categories are in line with previous reports and with general knowledge on horizontal transfer/duplication dynamics in bacteria (4,5,39). Metabolic genes are significantly enriched with Poisson-like abundance profiles, and in negative association with peaked families. This supports the idea that these domain (super)families are mainly sculpted by horizontal transfers. The gene-regulation category is significantly associated with overdispersed families, and underrepresented in peaked families, and intracellular processes follow the same trend. Finally, the information functional category is significantly associated with peaked families, and strongly underrepresented in overdispersed and Poisson families, suggesting that families related to translation should be subject to abundance constraints, absent in the null model. A parallel analysis that includes the noisy classes yields results that are very consistent (Supplementary Table S1).

The analysis of the finer functional categories is presented in Supplementary Table S2. Although affected by small set statistics, this analysis contains some further information. As expected, most of the signal from the information functional category comes from families associated with translation. For metabolism-associated families,

Table 1. Relation of the abundance profile of a family with its biological function

	Total	Poisson	Overdispersed	Peaked	Zero Variance
	701	282	130	281	8
General	26	11	6	9	0
Information	118	21 ($\mathbf{1.2 \cdot 10^{-8}}$)	10 ($\mathbf{8.1 \cdot 10^{-4}}$)	85 ($\mathbf{1.2 \cdot 10^{-14}}$)	2
Metabolism	318	151 ($\mathbf{2.4 \cdot 10^{-4}}$)	53	112 ($\mathbf{1.0 \cdot 10^{-2}}$)	2
Processes EC	17	8	3	5	1
Processes IC	85	35	27 ($\mathbf{1.2 \cdot 10^{-3}}$)	23 ($\mathbf{5.5 \cdot 10^{-3}}$)	0
Regulation	66	29	20 ($\mathbf{1.0 \cdot 10^{-2}}$)	16 ($\mathbf{3.6 \cdot 10^{-3}}$)	1
Other	41	16	8	15	2
Not Annotated	30	11	3	16	0

The counts for the larger functional categories of domain superfamilies divided according to their abundance profile histograms following thresholds on Q_f (see text) are reported. *P*-values for Fisher's exact tests are reported in parenthesis when significant ($P < 0.01$), in green for overrepresentation and in red for underrepresentation. Only the 701 families below the noise threshold for the family abundance histograms (see text) are considered. A wider classification yields similar results (Supplementary Table S1). A similar analysis with the finer functional categories is presented in Supplementary Table S2.

the tendency to have a Poisson-like abundance histogram is particularly strong for domain topologies associated with oxidation/reduction. In the regulation functional category, DNA-binding/transcription factor families are associated with overdispersed abundance histograms. A similar trend is found for signal transduction (such as, e.g. the GAF or EAL domains), while kinase/phosphatase domain families are more strongly associated with Poisson-like abundance profiles. Within the intracellular processes category, families associated with proteases and transport (e.g. porins) also follow a similar trend, while protein modification (chaperones) is enriched for peaked family abundance profiles. Finally, within the extracellular processes category, cell adhesion families are significantly associated with overdispersed profiles. Interestingly, transposases, which are a main driver of mobile elements in bacterial genome, do not show any specific pattern in abundance fluctuations, and different transposase superfamilies are found to have a wide range of values of the index Q_f.

DISCUSSION

The topic of which families and classes of genes can be more frequently transferred than others has been thoroughly studied in recent years (40–43). This study takes a complementary approach and focuses on the inter-species family abundance statistics, offering a novel observable, the abundance fluctuations, which our analysis connects to family-specific HGT dynamics. Importantly, abundance fluctuations are nontrivially distinct from family occurrence (average presence across species), which is the equivalent for domain families of the gene-frequency distribution (16–20). Indeed, the occurrence of a family is not a good predictor of the abundance fluctuation index Q_f (Figure 3B). Hence, abundance fluctuations are not recapitulated by occurrence, which appears natural from a biological standpoint, as adding or removing an entire family from a genome is different, in terms of accessible functional landscape, than adding or removing single family members.

The null model, which we use as a tool to rationalize the link between abundance and HGT dynamics, is based on the following assumptions. The main (null) assumption is that inter-species gene-family abundance fluctuations corre-

spond, effectively, to a stochastic process that realizes a minimal description of inter-species HGTs, intra-species family expansion (by duplication and HGT), and gene losses occurring in a family (7,44). The model focuses on species and does not include population dynamics (45,46); this approach has proven successful in other contexts (9–11,33,47). Each event occurring in one species is to be interpreted as the fixation for the gain or loss of a gene of a particular family in the species representative genome. A final simplifying assumption describes different gene families as independent random variables. It is likely that the abundance of different gene classes depends on biological interactions between them (32). Here, family interactions are described effectively as family-specific effective rates of fixed events. Differently from other descriptions, where horizontal transfer dynamics is treated as an effective interaction with some background gene pool (11,18,20,33), our minimal model describes cross-species horizontal transfers by explicitly accounting for interactions between genomes. The resulting 'collisional' formalism has a parallel with the framework classically used by Boltzmann in statistical physics (26). While this has no deep scientific meaning, it has some technical significance: it allowed us to make use of existing analytical techniques to directly access and control the model (48).

One further observation that should be made is that the model assumption of pairwise random interaction between genomes is rather primitive. The assumption is equivalent to considering the genomes equally related to each other (i.e. no genome-level tree exists). This is a gross simplification that is only partially mitigated in our data analysis by using a species-representative set of genomes. Instead, inter-species HGTs occur with decreasing probability for increasing evolutionary distance between genomes (49). The standard formulation of the model does not include this bias, and assumes equal probability of collisions for all genomes. For this reason, we considered a model variant where the collisions are biased by the phylogenetic distance between species (less probable for increasingly distant species). Our simulations (Supplementary Figure S8) show that the results are robust if a bias in the probability of horizontal exchanges is introduced. We estimated this bias from the

phylogenetic distances of empirical genomes in our data set, scored by the occurrence patterns of domain families (37).

We now discuss the functional enrichment of families with different abundance profiles. Three main patterns appear. Poisson-like abundance profiles correspond to families dominated by HGT, enriched for metabolic enzymes and kinases/phophatases. Families with overdispersed abundance profiles are characterized by an interplay of HGT and duplications (or intra-species HGT), and enriched for DNA-binding/transcription factors, signal transduction, and functional categories related intracellular processes such as transport and proteases. Families with peaked abundance profiles typically show no HGTs and are highly enriched for translation-related functions. In more detail, metabolic enzymes typically have Poisson-like abundance profile, and thus should be dominated by HGT with respect to duplications. Transferred metabolic enzymes are reported to outnumber duplicated enzymes in the recent evolution of *E. coli*, determining adaptation to new environments (12). Conversely, transcription factors are associated with overdispersed abundance profiles, which, in the model, correspond to increased gene duplications or intra-species horizontal transfers. Analyses of transcriptional regulatory networks (50,51) have found an enrichment of interactions involving paralogs, which can be explained by regulatory divergence after gene duplication. However, since it is well known that horizontal transfer should be the main drive of bacterial genome innovation, this observation has generated some debate (52,53). A recent observation that might reconcile the debate is that widespread intra-species transfers of regulatory regions could explain the regulatory divergence (54). Since the model focuses on inter-species events, it is compatible with this hypothesis. Indeed, as mentioned above, the model does not distinguish duplications from intra-species transfers. A recent analysis indicates that the amount of transfers is proportional to family size (35); it is presumable that horizontal transfers may effectively follow a rich-gets-richer principle, similarly to gene duplications (55), in line with the model assumptions. Regarding the data analysis, the large-scale data available to us only scored transfers, which are frequent for families with overdispersed abundance profiles. A finer, smaller-scale analysis also including duplications is compatible with our results (Supplementary Figure S7), but does not allow to formulate any specific hypothesis on the relative roles of intra-species transfers and duplications in families with overdispersed abundance profiles.

Families with peaked abundance profiles or zero variance are not captured by the model, and hence falsify its null assumptions. Since their abundance varies less than expected across genomes, we speculate that these families might be subject to adaptive abundance constraints, i.e. the domains are needed only in a certain number of 'copies', probably corresponding to properties of the protein machines that they form. Notably, while the majority of the peaked abundance profiles correspond to families where little or no horizontal transfers are reported, there are peaked families with high HGT: we have considered the set of top 15% H_f families for both the HGT-DB and DarkHorse data sets, and we found 37 and 35 peaked families, respectively. These lists are mildly enriched only for the metabolism functional category

(21/37, $P = 0.01$ for HGT-DB data and 18/35, $P = 0.2$ for DarkHorse; the full lists are given as supplementary files), while the other categories (including translation, which is enriched in peaked families) appear sparsely. The domains common to both lists include two domains for the metabolic enzyme cysteine methyltransferase, involved in DNA repair, the ligase domain ADC synthase, and a gated mechanosensitive ion channel domain. Notably, five distinct ATP synthase domains are found in the HGT-DB list of peaked families with high H_f; this gene has a very unusual phylogenetic distribution indicative of HGT (56). Our observations lead to hypothesize the presence of a (to our knowledge so far unreported) constraint on the abundance, possibly of selective origin. A more detailed investigation of this issue might be worth the effort.

Our analysis indicates that the abundance profile of a family generates an informed hypothesis about its function and evolution. More importantly, this suggests a 'segmented', or modular view of the genome (57), where the inter-species diversity of a gene family is highly dependent on its biological function. These considerations are in line with those emerging from pan- and core-genome analyses (58), but considering protein abundance fluctuations gives a different perspective than gene presence/absence patterns. For example, we have shown evidence that gene families on a genome with different abundance variability will also have different plasticity (or 'fluidity') (14,59), determined by the rates of the different evolutionary processes. As a result, some gene families will see the phylogenetic tree as a tree, while some others will perceive it as a network (whose effective topology can vary from family to family) (5,60).

CONCLUSION

The results show that cross-genomic family abundance fluctuations give access to relevant information regarding the processes by which gene families expand and contract in genomes and on the exchange of genes between genomes. We introduced a novel modeling framework to enhance and guide the data analysis. The model can help rationalize the most important finding of this work, which establishes a link between cross-species abundance fluctuations and HGT dynamics. The joint analysis of model and data suggests a classification of families based on their cross-genomic abundance histograms, which is biologically meaningful, as the different types of abundance profiles correspond well with the reasonable expectations in terms of biological functions. Importantly, while the necessarily simplifying model assumptions can be discussed, the resulting classification of the data is fully coherent for genomes with different size ranges and finds direct support from cross-comparisons with independent data sources. This suggests that its validity goes well beyond the necessarily radical simplifications of the model. Finally, we showed how abundance fluctuations provide a new perspective on genome plasticity. We speculate that a more detailed quantification of plasticity variation across families will be important in future studies to assemble a unified evolutionary view of genome architecture.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

REFERENCES

1. Shapiro,J.A. (2013) How life changes itself: the Read-Write (RW) genome. *Phys. Life Rev.*, **10**, 287–323.
2. Doolittle,R.F. and Bork,P. (1993) Evolutionarily mobile modules in proteins. *Sci. Am.*, **269**, 50–56.
3. Orengo,C.A. and Thornton,J.M. (2005) Protein families and their evolution—a structural perspective. *Annu. Rev. Biochem.*, **74**, 867–900.
4. Koonin,E.V. and Wolf,Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, **36**, 6688–6719.
5. Gogarten,J.P. and Townsend,J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, **3**, 679–687.
6. Huynen,M. and van Nimwegen,E. (1998) The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.*, **15**, 583–589.
7. Koonin,E.V. (2011) Are there laws of genome evolution?. *PLoS Comput. Biol.*, **7**, e1002173.
8. Boussau,B. and Daubin,V. (2010) Genomes as documents of evolutionary history. *Trends Ecol. Evol.*, **25**, 224–232.
9. Qian,J., Luscombe,N.M. and Gerstein,M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.*, **313**, 673–681.
10. Karev,G.P., Wolf,Y.I., Berezovskaya,F.S. and Koonin,E.V. (2004) Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol. Biol.*, **4**, 32.
11. Cosentino Lagomarsino,M., Sellerio,A., Heijning,P. and Bassetti,B. (2009) Universal features in the genome-level evolution of protein domains. *Genome Biol.*, **10**, R12.
12. Pal,C., Papp,B. and Lercher,M.J. (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.*, **37**, 1372–1375.
13. Treangen,T.J. and Rocha,E.P.C. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.*, **7**, e1001284.
14. Kislyuk,A.O., Haegeman,B., Bergman,N.H. and Weitz,J.S. (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics*, **12**, 32.
15. Andersson,D. and Hughes,D. (2009) Gene amplification and adaptive evolution in bacteria. *Annu. Rev. Genetics*, **43**, 167–295.
16. Lobkovsky,A.E., Wolf,Y.I. and Koonin,E.V. (2013) Gene frequency distributions reject a neutral model of genome evolution. *Genome Biol. Evol.*, **5**, 233–242.
17. Pang,T.Y. and Maslov,S. (2013) Universal distribution of component frequencies in biological and technological systems. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 6235–6239.
18. Haegeman,B. and Weitz,J.S. (2012) A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics*, **13**, 196.
19. Collins,R.E. and Higgs,P.G. (2012) Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Mol. Biol. Evol.*, **29**, 3413–3425.
20. Baumdicker,F., Hess,W.R. and Pfaffelhuber,P. (2012) The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.*, **4**, 443–456.
21. Wilson,D., Pethica,R., Zhou,Y., Talbot,C., Vogel,C., Madera,M., Chothia,C. and Gough,J. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.
22. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res.*, **32**, 226–229.
23. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, 308–313.
24. Garcia-Vallv,S., Guzmn,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.
25. Podell,S., Gaasterland,T. and Allen,E. (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformatics*, **9**, 419.
26. Cercignani,C. (1988) *The Boltzmann equation*, Springer, New York.
27. Ispolatov,S., Krapivsky,P. and Redner,S. (1998) Wealth distributions in asset exchange models. *Eur. Phys. J. B-Condens. Matter Complex Syst.*, **2**, 267–276.
28. Matthes,D. and Toscani,G. (2008) On steady distributions of kinetic models of conservative economies. *J. Stat. Phys.*, **130**, 1087–1117.
29. Chatterjee,A. and Chakrabarti,B.K. (2007) Kinetic exchange models for income and wealth distributions. *Eur. Phys. J. B*, **60**, 135–149.
30. Lallouache,M., Chakrabarti,A.S., Chakrabarti,A. and Chakrabarti,B.K. (2010) Opinion formation in kinetic exchange models: spontaneous symmetry-breaking transition. *Phys. Rev. E*, **82**, 056112.
31. Villani,C. (1998) On a new class of weak solutions to the spatially homogeneous Boltzmann and Landau equations. *Arch. Ration. Mech. Anal.*, **143**, 273–307.
32. Grilli,J., Bassetti,B., Maslov,S. and Cosentino Lagomarsino,M. (2012) Joint scaling laws in functional and evolutionary categories in prokaryotic genomes. *Nucleic Acids Res.*, **40**, 530–540.
33. Molina,N. and van Nimwegen,E. (2008) The evolution of domain-content in bacterial genomes. *Biol. Direct*, **3**, 51.
34. van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.*, **19**, 479–484.
35. Grassi,L., Caselle,M., Lercher,M.J. and Cosentino Lagomarsino,M. (2012) Horizontal gene transfers as metagenomic gene duplications. *Mol. Biosyst.*, **8**, 790–795.
36. Podell,S. and Gaasterland,T. *et al.* (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.*, **8**, R16.
37. Abeln,S. and Deane,C.M. (2005) Fold usage on genomes and protein fold evolution. *Proteins*, **60**, 690–700.
38. Lang,J.M., Darling,A.E. and Eisen,J.A. (2013) Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS One*, **8**, e62510.
39. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
40. Cohen,O. and Pupko,T. (2010) Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol. Biol. Evol.*, **27**, 703–713.
41. Ragan,M.A. and Beiko,R.G. (2009) Lateral genetic transfer: open issues. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **364**, 2241–2251.
42. Sorek,R., Zhu,Y., Creevey,C.J., Francino,M.P., Bork,P. and Rubin,E.M. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science*, **318**, 1449–1452.
43. Koonin,E.V., Makarova,K.S. and Aravind,L. (2001) Horizontal Gene Transfer In Prokaryotes: Quantification and Classification1. *Annu. Rev. Microbiol.*, **55**, 709–742.
44. Lynch,M. (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U.S.A.*, **104**(Suppl. 1), 8597–8604.
45. Novozhilov,A.S., Karev,G.P. and Koonin,E.V. (2005) Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.*, **22**, 1721–1732.
46. Berg,O.G. and Kurland,C.G. (2002) Evolution of microbial genomes: sequence acquisition and loss. *Mol. Biol. Evol.*, **19**, 2265–2276.

47. Maslov,S., Krishna,S., Pang,T.Y. and Sneppen,K. (2009) Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9743–9748.

48. Bassetti,F., Ladelli,L. and Matthes,D. (2011) Central limit theorem for a class of one-dimensional kinetic equations. *Probab. Theory Relate. Fields*, **150**, 77–109.

49. Andam,C.P. and Gogarten,J.P. (2011) Biased gene transfer in microbial evolution. *Nat. Rev. Microbiol.*, **9**, 543–555.

50. Teichmann,S.A. and Babu,M.M. (2004) Gene regulatory network growth by duplication. *Nat. Genet.*, **36**, 492–496.

51. Cosentino Lagomarsino,M., Jona,P., Bassetti,B. and Isambert,H. (2007) Hierarchy and feedback in the evolution of the Escherichia coli transcription network. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 5516–5520.

52. Price,M.N., Dehal,P.S. and Arkin,A.P. (2008) Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. *Genome Biol.*, **9**, R4.

53. Martnez-Nez,M.A., Prez-Rueda,E., Gutirrez-Ros,R.M. and Merino,E. (2010) New insights into the regulatory networks of paralogous genes in bacteria. *Microbiology*, **156**, 14–22.

54. Skippington,E. and Ragan,M.A. (2011) Within-species lateral genetic transfer and the evolution of transcriptional regulation in Escherichia coli and Shigella. *BMC Genomics*, **12**, 532.

55. Grassi,L., Grilli,J. and Cosentino Lagomarsino,M. (2012) Large-scale dynamics of horizontal transfers. *Mob. Genet. Elem.*, **2**, 163–167.

56. Hilario,E. and Gogarten,J.P. (1993) Horizontal transfer of ATPase genes—the tree of life becomes a net of life. *Biosystems*, **31**, 111–119.

57. Rivera,M.C., Jain,R., Moore,J.E. and Lake,J.A. (1998) Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 6239–6244.

58. Lapierre,P. and Gogarten,J.P. (2009) Estimating the size of the bacterial pan-genome. *Trends Genet.*, **25**, 107–110.

59. Toussaint,A. and Chandler,M. (2012) Prokaryote genome fluidity: toward a system approach of the mobilome. *Methods Mol. Biol.*, **804**, 57–80.

60. Philippe,H. and Douady,C.J. (2003) Horizontal gene transfer and phylogenetics. *Curr. Opin. Microbiol.*, **6**, 498–505.