Published online 9 June 2017

Nucleic Acids Research, 2017, Vol. 45, No. 13 ***7615–7622***
doi: 10.1093/nar/gkx510

Family-specific scaling laws in bacterial genomes

Eleonora De Lazzari[1], Jacopo Grilli[2], Sergei Maslov[3,*] and Marco Cosentino Lagomarsino[1,4,5,*]

[1]Sorbonne Universités, UPMC Université Paris 06, UMR 7238 Computational and Quantitative Biology, Genomic Physics Group, 4 Place Jussieu, Paris 75005, France, [2]Department of Ecology and Evolution, University of Chicago, 1101 E 57th st 60637 Chicago, IL, USA, [3]Department of Bioengineering, Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, [4]CNRS, UMR 7238, Paris, France and [5]FIRC Institute of Molecular Oncology (IFOM), 20139 Milan, Italy

Received March 07, 2017; Revised May 29, 2017; Editorial Decision May 30, 2017; Accepted May 30, 2017

ABSTRACT

Among several quantitative invariants found in evolutionary genomics, one of the most striking is the scaling of the overall abundance of proteins, or protein domains, sharing a specific functional annotation across genomes of given size. The size of these functional categories change, on average, as power-laws in the total number of protein-coding genes. Here, we show that such regularities are not restricted to the overall behavior of high-level functional categories, but also exist systematically at the level of single evolutionary families of protein domains. Specifically, the number of proteins within each family follows family-specific scaling laws with genome size. Functionally similar sets of families tend to follow similar scaling laws, but this is not always the case. To understand this systematically, we provide a comprehensive classification of families based on their scaling properties. Additionally, we develop a quantitative score for the heterogeneity of the scaling of families belonging to a given category or predefined group. Under the common reasonable assumption that selection is driven solely or mainly by biological function, these findings point to fine-tuned and interdependent functional roles of specific protein domains, beyond our current functional annotations. This analysis provides a deeper view on the links between evolutionary expansion of protein families and the functional constraints shaping the gene repertoire of bacterial genomes.

INTRODUCTION

As demonstrated by van Nimwegen (1) and confirmed by a series of follow-up studies (2–6), striking quantitative laws exist for high-level functional categories of genes. Specifically, the number of genes within individual functional categories such as e.g. that of transcriptional regulators (1,7,8) exhibit clear power-laws, when plotted as a function of genome size measured in terms of its number of protein-coding genes or, at a finer level of resolution, of their constitutive domains. In prokaryotes, such scaling laws appear well conserved across clades and lifestyles (9), supporting the simple hypothesis that these scaling laws are universally shared by this group.

From the evolutionary genomics viewpoint (10), these laws have been explained as a byproduct of specific 'evolutionary potentials', i.e. per-category-member rates of additions/deletions fixed in the population over evolution. As predicted by quantitative arguments, estimates of such rates correlate well with the category scaling exponents (1,2). A complementary point of view (5,8,11) focuses on the existence of universal 'recipes' determining ratios of proteins between different functions. Such recipes should mirror the 'dependency structure' or network operating within genomes as well as other complex systems (12). According to this point of view the usefulness, and thus the occurence, of a given functional component depends on the presence of a set of other components, which are necessary for it to be operational.

Beyond functional categories, protein coding genes can be classified in 'evolutionary families' defined by the homology of their sequences. Functional categories routinely contain genes from tens or more of distinct evolutionary families. The statistics of gene families also exhibits quantitative laws and regularities starting from a universal distribution of their per-genome abundance (13), explained by evolutionary models accounting for birth, death, and expansion of individual families (14–16).

While some earlier work connects per-genome abundance statistics of families with functional scaling laws (5), the link between functional category scaling and evolutionary expansion of gene families that build them remains rela-

*To whom correspondence should be addressed. Tel: +33 144277341; Email: marco.cosentino-lagomarsino@upmc.fr
Correspondence may also be addressed to Sergei Maslov. Tel: +1 217 265 5705; Email: maslov@illinois.etu

tively unexplored. Clearly, selective pressure is driven by functional constraints, and thus selection cannot in principle recognize families with identical functional roles. On the other hand, slight differences in the functional spectrum of different protein domains, and interdependency of different functions can make the scenario more complex. Thus, one central question is how the abundance of genes performing a specific function emerges from the evolutionary dynamics at the family level.

Two alternative extreme scenarios can be put forward: (i) the high-level scaling laws could emerge only at the level of functions, and be 'combinatorially neutral' at the level of the evolutionary families building up a particular function, or, *vice versa*, (ii) they could be the result of the sum family-specific scaling laws. In the first scenario all or most of the families performing a particular function would be mutually interchangeable. In the second scenario, the evolutionary potentials would be family specific and coincide with family evolutionary expansion rates, possibly emerging from the complex dependency structure cited above and from fine-tuned functional specificity of distinct families. An intermediate possibility is that an interplay of constraints acts on both functional and evolutionary families. The first test for the feasibility of the second scenario is the existence of scaling laws for individual families. Here, focusing on bacteria, and using protein domains to define families, we present a clear evidence for family-specific scaling laws with genome size. We show that the abundance of the families follows power laws with genome size. Comparing functional categories with a suitable null model, we show that family-specific exponents may deviate significantly from the exponent of the associated functional category. We provide a comprehensive classification of families based on common scaling exponents, which recovers the known functional associations as well as revealing new ones, and may be used to detect possible misannotations. Finally, we develop quantitative tools to measure the heterogeneity of the scaling of families belonging to a given category or predefined group of families.

MATERIALS AND METHODS

Data sources

We considered bacterial proteomes retrieved from the SUPERFAMILY (release 1.75 downloaded in October 2014, (17)) and PFAM (release 27.0 downloaded in October 2014, (18,19)) database. Evolutionary families were defined from the domain assignments of 1535 superfamilies (SUPERFAMILY database) and 446 clans (PFAM database) on all protein sequences in completed genomes. We focused the analysis on the 1112 bacterial proteomes used as species reference in the SUPERFAMILY database. For the functional annotations of the SUPERFAMILY data, we considered annotation of SCOP domains as a scheme of 50 more detailed functional categories, mapped to 7 more general function categories, developed by C. Vogel (20). PFAM clans were annotated on the same scheme of 50 functional categories, using the mapping of clans into superfamilies available from the PFAM website http://pfam.xfam.org/clan/browse#numbers (21).

Data analysis

For each evolutionary domain family (or a functional category consisting of multiple evolutionary families), genome sizes (measured in the overall number of domains) were logarithmically binned. For each bin we calculated mean and standard deviation of the given family abundance (number of domains) within the bin. The estimated scaling exponent β_i for family i is the result of the non-linear least squares fitting of the binned data weighted by the standard error of family abundance. Genome size bins containing <10 genomes were not taken into account. To filter out the data that, due to low-abundance or rare families, were affected by sampling problems, we considered three independent parameters, (i) the 'occurrence', i.e. the fraction of genomes where family i is present, $o_i = N_G^{(i)}/N_G$, where N_G is the total number of genomes in the sample, and $N_G^{(i)}$ is the number of genomes where the family has non-zero abundance, (ii) the goodness of fit index

$$s_i = \frac{1}{1+\sqrt{LS_i}}$$

where, LS_i is the error associated with the exponent β_i, measured as the average squared deviation between the fit and the logarithm of the empirical abundance (see Supplementary Note S1) and (iii) the Pearson correlation coefficient ρ_i between the logarithm of the family abundance and the logarithm of the genome size. The index s_i puts on the same ground families with different exponents, but generally decreases as the scaling exponent increases, in accordance with the growth of fluctuations in families with higher exponents observed in ref. (22). Hence, we decided to use it only for low exponents, where the Pearson correlation is a bad proxy of scaling. We considered families with $s_i > 0.9$ and $o_i > 0.6$ for exponents <0.2, otherwise families with $\rho_i > 0.4$ and $o_i > 0.6$ reducing the dataset to 357 superfamilies and 178 clans that satisfy both requirements. As shown in Supplementary Figure S1A, s_i and o_i are not mutually correlated across the genomes, implying that the two requirements are in fact independent, the same is valid for ρ_i and o_i, see Supplementary Figure S1B. We verified that the removed families with the procedure described above do not influence the scaling of the category. Supplementary Figure S2 reports the exponent of the category scaling before the thresholding (where all the families are considered) and after (where the domains belonging to the removed families are not considered in the category scaling), showing that the values are consistent for all the categories studied.

For each family within a given functional category, we defined a 'heterogeneity score' h_i as follows:

$$h_i = |\beta_c - \beta_i|,$$

where, β_i and β_c are,respectively, the scaling exponents of family i and functional category c. The heterogeneity measure for each functional category was defined as the average of the per-family heterogeneity scores h_i:

$$H_c = \frac{1}{F_c}\sum_i h_i,$$

where, F_c is the number of families in category c.

The significance of the values found with this formula was assessed against a null model assuming that the total abundance of a category is distributed randomly across the associated families. The average abundance (i.e. the fraction of domains belonging to a family averaged over genomes) and occurrence (fraction of genomes where the family is present) of each family are both conserved (note that these two properties are uncorrelated in the data, hence we chose to conserve both in the null model, assuming that they are independent, see Supplementary Figure S3).

Given a genome g with n_c^g elements in the functional category c, divided into F_c^g associated families, we redistributed the n_c^g members among the F_c^g sets conserving the average relative abundance of each family (see Supplementary Note S2). A member of family i belonging to category c was therefore added with probability:

$$p_{i,c} \propto \begin{cases} \dfrac{1}{N_g^{(i)}} \displaystyle\sum_{g'=1}^{N_G^{(i)}} \dfrac{n_i^{g'}}{\sum_{k\in c} n_k^{g'}} & , \text{ if } n_i^g \neq 0 \\ 0 & , \text{ if } n_i^g = 0 \, . \end{cases}$$

The resulting set of F_c artificially built evolutionary categories constrains the occurrence pattern and the average abundance of the original ones. Scaling exponents for families in the null model are extracted with the procedure described above. Only functional categories containing domains from more than 10 distinct families were compared to the null model. All procedures were implemented as custom Python 2.7 scripts.

RESULTS

Families have individual scaling exponents, reflected by family-specific scaling laws

We started by addressing the question of whether individual families show scaling laws, and thus can be associated to specific scaling exponents. In order to do so, we isolated domains belonging to the same family across the sample of 1112 species-representative bacterial proteomes and plotted their abundance against the total number of domains in the corresponding proteome.

When the abundance is sufficiently high to overcome sampling problems, most families show a clearly identifiable individual scaling when plotted as a function of genome size. As an example, Figure 1 shows the scaling of a set of chosen families in four selected functional categories. Additionally, some low-abundance families that occur in all genomes with a very consistent number of copies show definite scaling with exponents close to zero (22), being clearly constant with size, with little or no fluctuations.

Given that functional categories follow specific scaling laws, likely related to function-specific evolutionary trends (1,2), there remain different open possibilities for the behavior of the evolutionary families composing the functional categories. One simple scenario is that family scalings are family specific, thus validating the existence of family evolutionary expansion rates that are quantitatively different to the one of their functional category. In the opposite extreme scenario the scaling is only function specific, and individual families performing similar functions are interchangeable.

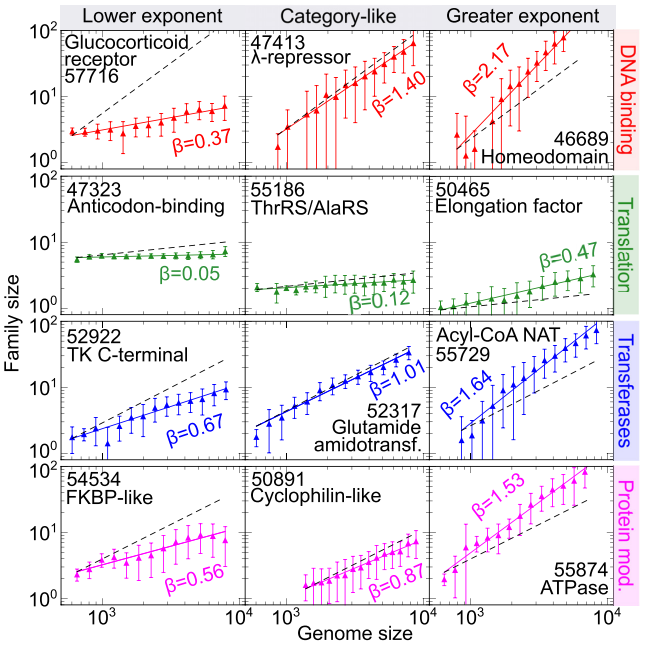

Figure 1. Families follow specific scaling laws, which may agree or deviate from the overall scaling of the functional category to which this family belongs. The plots report the abundance of 12 different superfamilies as a function of the genome size (triangles are binned averages). The power-law fits (solid lines) are compared to the power-law fits of the functional category to which each family belong (dashed black lines). We display here examples from four functional categories: DNA binding (top row), Translation (second row from top), Transferases (third row from top) and Protein modification (bottom row). Families in the leftmost/rightmost column scale respectively slower/faster than their category means, families in the middle column have similar slope to the full category. Legends specify the SCOP superfamily id, family descriptive name and power-law exponent (β_i) from the fits. Scaling lines for functional categories were shifted vertically in order to intersect empirical scaling data for families at the leftmost point. The original intercepts are: 0.0007 ± 0.0143 (DNA binding), 39.131 ± 0.006 (Translation), 0.04 ± 0.02 (Transferases) and 0.01 ± 0.03 (Protein modification).

If this were the case, family diversity in scaling exponent would be only due to sampling effects, and the null model would fully reproduce the diversity in family scaling observed in empirical data. To address this question, we randomized the families within a category conserving their occurrence patterns and the category average abundance. The randomized families always show very similar scaling as the one of the corresponding category (see Supplementary Figure S4). Hence, this analysis strongly supports the existence of family-specific scaling exponents that do not simply descend from the category scaling.

Figure 1 shows that the presence of 'outlier families' is common among functional categories. In most categories, we found families where the deviations from the category exponents is clear, beyond the uncertainty due to the errors from the fits. Figure 1 shows some examples where in each of the shown categories β_i may be higher, lower or comparable to β_c. A table containing all the family and category exponents is available as supplementary information (Supplementary Tables S1 and 2).

Finally, we considered the correlation of family scaling exponents with relevant biological and evolutionary param-

Table 1. Family scaling exponents can be associated to specific biological functions

Detailed function	$\beta_i \leq 0.6$	$0.6 < \beta_i < 1.4$	$\beta_i \geq 1.4$	$\beta_c \pm \sigma_{\beta_c}$
Translation	20(4.3)	1(−3.7)	0	0.16±0.03
DNA replication/repair	11	7	0	0.51±0.07
Transport	5	9	1	1.1±0.2
Proteases	7	9	0	0.9±0.1
Protein modification	8	1(−2.3)	2	1.06±0.09
Ion m/tr	11	3	3(−2.2)	1.3±0.1
Other enzymes	29	32	2	1.04±0.06
Coenzyme m/tr	17(2.2)	6	1	0.85±0.09
Redox	4(−3.3)	18(3.1)	2	1.2±0.1
Energy	11	7	0	0.86±0.09
Nucleotide m/tr	16(3.1)	3(−2.5)	0	0.53±0.08
Carbohydrate m/tr	4	8	0	1.0±0.2
Transferases	5	11	1	1.05±0.07
Amino acids m/tr	7	6	0	0.8±0.2
DNA-binding	5	4	4(3.3)	1.5±0.1
Signal transduction	1(−2.7)	5	5(5.0)	1.6±0.2
Unknown function	9	7	0	0.98±0.09

Each cell in the table indicates the number of families that functional categories (rows) share with groups of families whose scaling exponents fall in pre-defined intervals (columns). The table also shows the *Z*-scores for a standard hypergeometric test (shown in green for over-representation and in red for under-representation, only $|Z| > 1.96$ are shown).

eters such as foldability (quantified by size-corrected contact order, SMCO (23)), the diversity of EC-numbers associated with families (quantifying the functional plasticity of a given family), selective pressure (quantified by the ratio of non-synonimous to synonymous K_a/K_s substitution rates (24)) and overall family abundance. The results are summarized in Supplementary Table S2. Foldability and K_a/K_s appear to have little correlation with scaling exponents. Instead, we found a significant positive correlation of exponents with family abundance, and both quantities are correlated with diversity of EC-numbers in metabolic families. This suggests that, at least for metabolism, functional properties of a fold play a role in family scaling, and that beyond metabolism, abundance and scaling are, on average, not unrelated.

The heterogeneity in scaling exponents is function specific

The analyses presented above support the hypothesis that functional categories contain families with specific scaling exponents. Supplementary Figure S5 reports a visual representation of the distributions of the family exponents within a category, complementing the information presented in Table 1. Indeed, the scaling exponents β_i of the families can be significantly different from the category exponent β_c, with deviations that are much larger than predicted by randomizing the categories according to the null model (see Supplementary Figure S4).

In order to quantify this 'scaling heterogeneity' of functional categories, we computed for each family *i* the distance between its scaling exponent β_i and the category exponent β_c (see 'Materials and Methods' section). We defined an index H_c quantifying the heterogeneity of the scaling of the families within a category by averaging this distance over the families associated to a given category *c*.

Figure 2A shows the relation between the heterogeneity H_c and the category exponent β_c. Interestingly, these two quantities are correlated, with categories with larger values of β_c being more heterogeneous. Intuitively, categories with small exponents are incompatible with extremely large fluctuations of family exponents, while categories with larger exponents can contain families with small β_i. Indeed, this trend of heterogeneity with exponents is also observed in the null model, where the heterogeneity of null categories is much smaller than empirical ones, since all families tend to take the exponent category (Supplementary Figure S4).

Figure 2B allows a direct comparison of the heterogeneity of different categories by subtracting the mean trend. It is noteworthy that the Signal Transduction functional category, which also has clear superlinear scaling, has much lower heterogeneity than DNA-binding/transcription factors. Among the categories with linear scaling, Transferases is one of the least heterogeneous ones, while the categories Protein Modification and Ion metabolism and Transport show a large variability in the exponents of the associated families. For Protein Modification, this signal is essentially due to the Gro-ES superfamily and to the HFSP90 ATP-ase domain, which have a clear superlinear scaling, while other chaperone families, such as FKBP, HSP20-like and J-domain are clearly sublinear with exponents close to zero. Interestingly, the Gro-EL domains, functionally associated to the Gro-EL, are part of this second class (exponent close to 0.2), showing very different abundance scaling to the Gro-EL partner domains. Conversely, the category Ion Metabolism and Transport is divided equally into linearly scaling (e.g. Ferritin-like Iron homeostasis domains) and markedly sublinear families, such as SUF (sulphur assimilation)/NIF (nitrogen fixation) domains. On the other hand, categories with small values of heterogeneity are made of families with exponents close to the one of the category, as shown in Table 1 in the case of, e.g. Transferases.

Note that, since the total abundance of a category is the sum of the abundances of the corresponding families, we expect that the narrower the distribution of family exponents within a category, the better the power-law approximation should hold at the functional categories level. Consequently, we tested the connection of category heterogeneity in exponents to goodness of fit. We found that the Spearman correlation coefficient between category heterogeneity and mean residual of the fit is equal to 0.43, indicating that more heterogeneous categories give slightly worse fits as expected by these considerations.

Determinants of the scaling exponent of a functional category

We have shown that scaling exponents of individual families may correspond to a variable extent to the exponent of the corresponding functional category. However, since categories are groups of families, the scaling of the former cannot be independent of the scaling of the latter. This section explores systematically the connection between the two. As detailed below, we find that in some cases the scaling exponent of functional categories is determined by few outlier

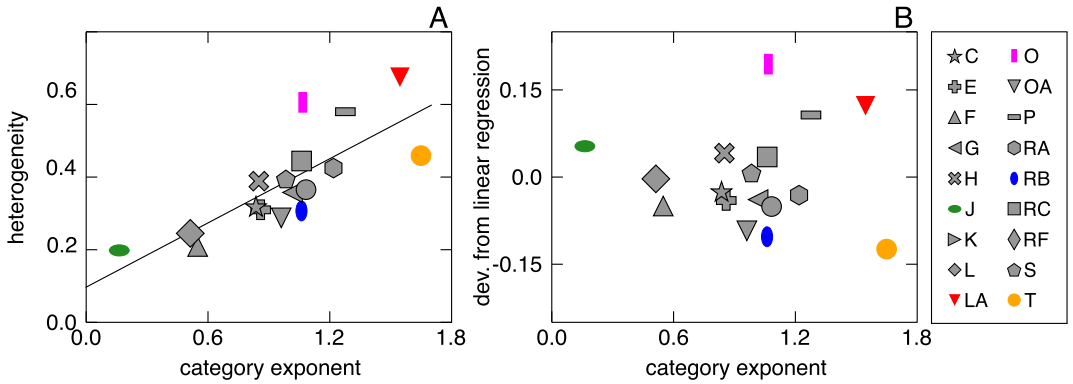

Figure 2. (**A**) Functional categories with faster scaling laws contain families with more heterogeneous scaling exponents. Heterogeneity is quantified by the mean deviation between the family scaling exponents and the category exponent. The plot reports heterogeneity scores for different functional categories, plotted as a function of the category exponents. The black line is the linear fit between heterogeneity and exponents (slope 0.3, intercept 0.1). (**B**) Comparison of heterogeneities subtracted from the linear trend. By this comparison, the least heterogeneous categories are Signal Transduction (T) and Transferase (RB), and the most heterogenous are DNA Binding (LA) and protein modification (O). Translation (J) is slightly above the trend for its low exponent. The legend (right panel) shows the association between symbols and category codes (see Supplementary Table S1 for the corresponding category name).

families, while in other cases most of the families within a category contribute to the category scaling exponent.

While many families have a clear power-law scaling, functional categories may contain many low-abundance families with unclear scaling properties. When considered individually, these families do not contribute much to the total number of domains of a category, but their joined effect on the scaling of the category could be potentially important. Supplementary Figure S6 shows that the sum of these low-abundance families does not suffer from sampling problems and shows a clear scaling. Interestingly, the scaling exponents for these sums once again does not necessarily coincide with the category exponents.

Figure 3A illustrates the systematic procedure that we used in order to understand how the scaling of categories emerges from the scaling of the associated families. Families were ranked by total abundance across all genomes (from the most to the least abundant) and removed one by one from the category. At each removal step in this procedure, both the scaling exponent of the removed family and the exponent of the remainder of the category are considered. In other words, the *i*-th step evaluates the exponent of the *i*-th ranking family (in order of overall abundance) and of the set of families obtained by removing the *i* top-ranking families (with highest abundance) from the category. The resulting exponents quantify the contribution of each family to the global category scaling, as well as the collective contribution of all the families with increasingly lower overall abundance.

The results (Figure 3B and Supplementary Figure S7), show how the heterogeneity features described above are related to family abundance. Pooled together, the low-abundance families within a functional category may show very different scaling than their category. Additionally, single families follow scaling laws that deviate from the one of the corresponding functional categories. One notable example of this are Transcription-Factor DNA-binding domains. If the abundance of the outliers families is large enough in terms of the fraction of domains in the functional category, they might be responsible for determining the scaling of the entire category, as it happens in the case of DNA-binding (which is more extensively discussed in the following section).

Overall, one can distinguish between two main behaviors, either a category scaling is driven by a low number of highly populated 'outlier' families (e.g. DNA binding and Protein Modification in Figure 3B), or the category scaling is coherent, and robust to family subtraction (e.g. Transferases and Translation in Figure 3B). While the first behavior appears to be more common for functional categories with higher scaling exponent, there are some exceptions. Notably, the scaling of strongly super-linear categories is not always driven by a few families. For example, the functional category Signal Transduction has an exponent $\beta_c = 1.7$, which remains stable after the removal of the largest families (Supplementary Figure S6 and Figure 3B). Both behaviors are clearly visible for intermediate exponents (in order to appreciate this, compare the Transferases and Protein Modification categories in Figure 3B).

Super-linear scaling of transcription factors is determined by the behavior of a few specific highly populated families

We considered, in particular, the case of DNA-binding/transcription factors (6), which are known to exhibit peculiar scaling in bacteria (8,25). The abundance of domains in this functional category increases superlinearly (almost quadratically) with the total domain counts (1,11,22). As shown in the first row of Figure 1B, not all the families in this functional category display a superlinear scaling (6), and the collective scaling of the low-abundance families with genome size is much slower (see Figure 3 and Supplementary Figure S6). Figure 3B shows that only the most five-six abundant families display a super-linear scaling ($\beta_i > 1$). These are Winged helix DNA-binding domains (34.8% of abundance), Homeodomain-like (23.3 %) lambda repressor-like DNA-binding domains (9.5%) bipartite Response regulators (7.7%) Periplasmic binding protein-like (6.2%) and FadR-like (2.4%). The remaining 16.1% of the DNA-binding regulatory domains follows a

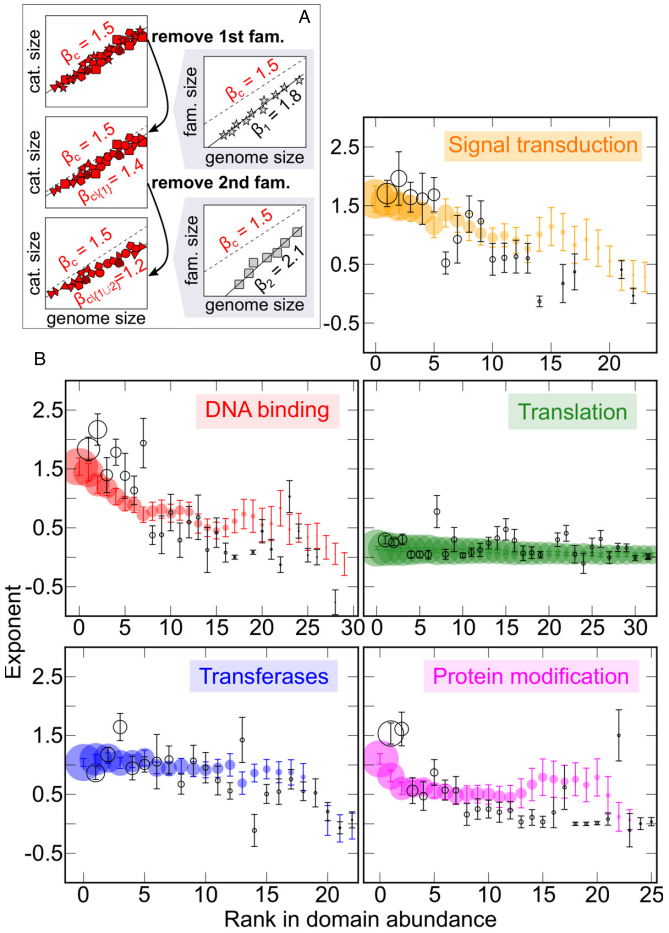

Figure 3. Systematic removal of families (ranked by abundance) inside functional categories reveals how individual families build up functional category scaling. (**A**) Illustration of the procedure. Families belonging to a given functional category are ranked by overall abundance on all genomes and removed one by one from the most abundant. The scaling of the removed family and the remainder of the category is evaluated after each removal. The plots are a stylized example of the first two steps (using values for the category DNA binding). β_c is the category exponent, β_i are family exponents and $\beta_{c\setminus\{i\}}$ are the stripped-category exponents, computed after the removals. (**B**) Results of this analysis for four functional categories. Empty circles represent the exponents β_i (and their errors) for the scaling law of each family belonging to the functional category (in order of rank in total abundance). Filled circles are the scaling exponents of functional categories without the domains of the i least abundant families. The size of each symbol is proportional to the fraction of domains in the family or family-stripped category. Error bars are uncertainties of the fits (see 'Materials and Methods' section). See Supplementary Figures S7 and 8 for the same plots obtained for other functional categories and using the PFAM database.

clear *sublinear* scaling with genome size (exponent 0.7, see Supplementary Figure S6).

Grouping families with similar scaling exponents shows known associations with biological function and reveals new ones.

The above analyses show that the range of scaling exponents of families within the same functional categories is generally wide and that the scaling behavior of some families sensibly deviates from their category. At the same time, functional categories show clear characteristic scaling laws, with well-defined exponents β_c (9). We, therefore, asked to what extent a range of family scaling exponents β_i is peculiar to a functional category and how this compares to the category exponent β_c. To this end, we grouped families based on their scaling exponents. We then used those groups to test how much specific range of exponents define specific functions by an enrichment test of functional annotations.

Table 1 shows that in most cases functional categories are over-represented in the exponent range where their scaling exponents β_c is found. This confirms and puts in a wider perspective the previously reported strong association between abundance scaling with size and functional annotation. As can be expected from previous results, the functional category Protein Modification is an exception: this category is under-represented in the linear region even though its category exponent is ∼1.06, since it contains two strongly superlinear families and a bulk of families with sublinear scaling. This strong heterogeneity in scaling exponents is also visible in Figure 3B.

The results of this analysis are not sensitive to the chosen intervals for the scaling exponents. In order to show this, we performed a more systematic enrichment analysis, using sliding windows of exponents of width 0.4, and step 0.1, and plotting the Z-score for the enrichment as a function of the representative family exponent for each window (Supplementary Figure S9). The maxima of this plot define a representative exponent for each functional category, and can be compared to the exponent β_c measured directly from the plot of category abundance versus genome size (see Supplementary Figure S10). Interestingly, this analysis also shows that in many cases a single functional category is enriched for multiple groups of families with well-defined exponents, as in the case of the Protein Modification category. The cases of Ion Metabolism and Transport (already discussed), Coenzyme Metabolism and Transport, Redox also show clear indications of enrichment for two or more exponent groups. For the category Coenzyme Metabolism and Transport this is due to the presence of a single abundant family with scaling exponent close to 2, the acyl-CoA dehydrogenase NM domain-like, whose functional annotation is still not well defined. In the case of Redox, the most abundant families (Thioredoxin-like, 4Fe-4S ferredoxins, Metallo-hydrolase/Oxydoreductase) scale linearly, but there is a wide range of families with exponents between 0.5 and 1, and once again two fairly abundant outlier families with superlinear scaling (Glyoxalase/Bleomycin resistance protein/Dioxygenase, and ALDH-like), both with a fairly wide range of functional annotations.

DISCUSSION AND CONCLUSION

Our results gather a critical mass of evidence in the direction of family-specific expansion rules for the families of protein domains found in a genome. Although previous work had focused on individual transcription factor families (6), finding in some cases some definite scaling, no attempts were made to address this question systematically. The scaling laws for domain families appear to be very robust, despite of the limited sampling of families compared to functional annotations (which are super-aggregates of families and hence

have by definition higher abundance). In particular, the results are consistent between the different classifications of families we tested (SUPERFAMILY and PFAM, see Supplementary Note S4).

One may wonder whether the fact that the genome size is the sum of the abundances of all families (and of all categories) imposes a constraint on the observed scalings. This question is related to whether the genome size is an actual driver of the scaling laws or there are other drivers, such as additional constraints connecting sizes, possibly function- or family specific that determine the scaling laws. If this second case were true, then the detection of the driving variables would reveal some aspects of the 'recipes' connecting different functions or categories to form a genome (5,8).

Overall, our results indicate that scaling laws are measurable at the family level, and, given the heterogeneous scaling of families with the same functional annotations, families are likely a more reliable description level for the scaling laws than functional annotations. The interpretation of these scaling laws is related to the evolutionary dynamics of family expansion by horizontal transfer or gene duplication, and gene loss (1,10,26). Scaling exponents are seen as 'evolutionary potentials' (2), is based on a model of function-specific (multiplicative) family expansion rates. Assuming this interpretation, then our result that these rates may be different for different domain families having the same functional annotation may seem puzzling. Clearly, selective pressure can only act at the functional level, and if 2-fold were functionally identical, there should be reasonably no advantage selecting one with respect to the other and doing so at different specific rates. For example, a transcription factor using one fold to bind DNA rather than another one should be indistinguishable from one using a different fold, provided binding specificity and regulatory action are the same.

In view of these considerations, we believe that our findings support a more complex scenario for the interplay between domain families and their functions. Specifically, we put forward two complementary rationalizations. The first is that functional annotations group together different domains whose abundance is linked in different ways to genome size because of their different biochemical and biological functional roles. Such differences may range from slight biochemical specificities of different folds to plain misannotations. This is possible, e.g. with enzymes, where the biochemical range of two different folds is generally different. This observation might be related to the positive correlation we found between the number of EC numbers corresponding to a metabolic domain and its scaling exponent. However, such interpretation might be less likely applicable to, e.g. transcription factor DNA-binding domains, where functional annotation is fairly straightforward (27), but different scaling behaviors with genome size are nevertheless found.

The second potential explanation assumes the point of view where scaling laws are the result of functional interdependency between different domain families (8,28), then correlated fluctuations around the mean of family pairs should carry memory of such dependency structures (12). More in detail, there may be specific dependencies connecting the relative proportions of domains with both different and equal functional annotations that are present in the same genome, which might determine the family-specific behavior (5). While further analysis is required to elucidate these trends, we believe that gaining knowledge on functional dependencies would be an important step to understand the functional design principles of genomes. It is not possible at this stage to distinguish between these two explanations, and we surmise that they may both be relevant to explain the data.

Of notable importance is the case of the superlinear scaling of transcription factors, which has created notable debate in the past (8,29). For the first time, we look into how this trend is subdivided between the different DNA binding domains (27). Our analysis indicates that the superlinear scaling is driven by the few most abundant superfamilies (mostly winged-helix, homeodomain, lambda repressor). However, the remaining 10–20% of the functional category gives a clear sublinear scaling with genome size, which emerges beyond any sampling problems. We speculate that these other regulatory DNA-binding domains may be functionally different or behave differently over evolutionary time scales. Hence, the scaling of transcription factors with size in bacteria is driven by a small set of domain families with scaling exponent close to two, which take up most of the abundance, but does not appear to be peculiar of *all* transcription factors. A 'toolbox' model considering the role of transcription factors as regulator of metabolic pathways and the finite universe of metabolic reactions (8,11) predicts scaling exponents close to two for transcription factor families. According to our results, such model should be applicable to the leading TF families. Interestingly, the heterogeneity in the behavior in transcription factor DNA-binding domains is much higher than that of the other notable superlinear functional category, signal transduction, where removal of the leading families does not significantly affect the observed scaling of abundance with genome size. Given the clarity and uniformity of the scaling exponent, we speculate that possibly a toolbox-like model may be applicable to understand the overall scaling of this category.

Other categories clearly contain multiple sets of families with coherent exponents or single outlier families. In some cases, two main groups of families with different scaling behavior clearly emerge, and higher observed scaling exponents may be related to a wider range of functional annotations. We propose that such easily detectable trends can be used to revise and refine functional annotations of protein domains. Such functional annotations are currently largely curated by humans, and based on subjective and/or biased criteria. The analysis of family scaling gives an additional objective test to define the coherence of the families that are annotated under the same function. While yet-to-be-developed automated inference methods based on our observations could serve this purpose, the quantitative scores defined here already provide useful information. The heterogeneity of a functional category is an indication of how likely that group of domain families follows a coherent expansion rate over evolution. The enrichment scores for sets of families with a given range of scaling exponent helps to pinpoint the sets of families within the functional category that expand coherently with genome size.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Erik van Nimwegen, Madan Babu, Otto Cordero and Purushottam Dixit for helpful discussions.

FUNDING

Indo-French Centre for the Promotion of Advanced Research (IFCPAR/CEFIPRA) [5103-3]. Funding for open access charge: IFCPAR/CEFIPRA [5103-3].
Conflict of interest statement. None declared.

REFERENCES

1. van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.*, **19**, 479–484.
2. Molina,N. and van Nimwegen,E. (2008) The evolution of domain-content in bacterial genomes. *Biol. Direct*, **3**, 51.
3. Molina,N. and van Nimwegen,E. (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet.*, **25**, 243–247.
4. Cordero,O.X. and Hogeweg,P. (2009) Regulome size in Prokaryotes: universality and lineage-specific variations. *Trends Genet.*, **25**, 285–286.
5. 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.
6. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
7. Stover,C., Pham,X., Erwin,A., Mizoguchi,S., Warrener,P., Hickey,M., Brinkman,F., Hufnagle,W., Kowalik,D., Lagrou,M. *et al.* (2000) Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, **406**, 959–964.
8. 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.
9. Molina,N. and van Nimwegen,E. (2009) Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet.*, **25**, 243–247.
10. Koonin,E.V. (2011) Are there laws of genome evolution?. *PLoS Comput. Biol.*, **7**, e1002173.
11. Pang,T.Y. and Maslov,S. (2011) A toolbox model of evolution of metabolic pathways on networks of arbitrary topology. *PLoS Comput. Biol.*, **7**, e1001137.
12. 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.
13. Huynen,M. and van Nimwegen,E. (1998) The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.*, **15**, 583–589.
14. 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.
15. 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.
16. 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.
17. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
18. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
19. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
20. Vogel,C. and Chothia,C. (2006) Protein family expansions and biological complexity. *PLoS Comput. Biol.*, **2**, e48.
21. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
22. Grilli,J., Romano,M., Bassetti,F. and Cosentino Lagomarsino,M. (2014) Cross-species gene-family fluctuations reveal the dynamics of horizontal transfers. *Nucleic Acids Res.*, **42**, 6850–6860.
23. Debes,C., Wang,M., Caetano-Anolles,G. and Graeter,F. (2013) Evolutionary optimization of protein folding. *PLoS Comput. Biol.*, **9**, e1002861.
24. Ndhlovu,A., Durand,P.M. and Hazelhurst,S. (2015) EvoDB: a database of evolutionary rate profiles, associated protein domains and phylogenetic trees for PFAM-A. *Database*, **2015**, bav065.
25. Ranea,J.A.G., Buchan,D.W.A., Thornton,J.M. and Orengo,C.A. (2004) Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.*, **336**, 871–887.
26. Grassi,L., Caselle,M., Lercher,M.J. and Cosentino Lagomarsino,M. (2012) Horizontal gene transfers as metagenomic gene duplications. *Mol. Biosyst.*, **8**, 790–795.
27. Madan Babu,M. and Teichmann,S. (2003) Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.*, **31**, 1234–1244.
28. Grassi,L., Grilli,J. and Cosentino Lagomarsino,M. (2012) Large-scale dynamics of horizontal transfers. *Mob. Genet. Elements*, **2**, 163–167.
29. Ranea,J.A.G., Grant,A., Thornton,J.M. and Orengo,C.A. (2005) Microeconomic principles explain an optimal genome size in bacteria. *Trends Genet.*, **21**, 21–25.