Supplementary Text and Figures for Grilli *et al.*

S1. DESCRIPTION OF THE MODEL AND BASIC MEAN-FIELD RESULTS

This section discusses in more detail the analytical derivation of the scaling for the main observables of model I and II using a mean-field approach.

Consider a joint partioning of elementary units (domains or genes) in functional and evolutionary categories, as illustrated in Figure 1 of the main text. The elementary units (in our case domains), belong to a single evolutionary family i, and every family i belongs to one and only one functional category c.

The generic stochastic growth model considered here defines how new units are introduced into the system. The model is specified by a set of basic rates. The basic set of rates is constitued by the probabilities p_i that a newly added unit belongs to a certain class i. More in detail, we define a probability p^i_O (where O stands for "old") that a new domain belongs to a family i which is already present in the system (i.e. having at least one member) and the probability p_N (where N stands for "new") that the added unit belongs to a family which is not already present in the system.

The choice of p^i_O and p_N defines the model as a stochastic process for the basic observables (such as genome size n, family number f and its population n_i, etc.), but one extra detail is needed. When a new class is introduced, the model needs to specify the category it belongs to. As discussed in the main text, in the model considered here a newly added family always belongs to a category c with probability χ_c. The probabilities p^i_O, p_N and χ_c can depend, in principle, on the number of units n and on their distribution in families, on the total number of families f and so on. Empirical data indicate (see Figure 2 in the main text) that χ_c is a category-dependent constant, and thus does not depend on n.

The mean-field approximation is useful to extract the basic information from the model [6]. In each realization of the full stochastic process, the probabilities of the possible configurations at time $t + 1$ are determined by the configuration at time t. The mean-field approximation assumes that the configuration at time t is the average configuration. For example, if one is interested in the number of domains belonging to family i, the average number of elements $n_i(t+1)$ at time $t+1$ will be equal to the average number of elements $n_i(t)$ at time t summed with the average number of elements added in a time step, i.e. p^i_O. For asymptotically large t this implies the approximate equation $\partial_t n_i = p_O^i$ for the averages (here the averaging procedure is implicit in the notation). Since typically, at each step one and only one element is added, the mean number of elements is $n = t$. If this is not the case, we can obtain $\partial_n n_i$ simply from $\partial_t n_i$ divided by $\partial_t n$. Considering $n = t$ we obtain, for a generic model, the following mean-field equations

$$\begin{aligned}
&\partial_n n_i = p_O^i \\
&\partial_n f = p_N \\
&\partial_n f_c = \chi_c p_N \\
&\partial_n n_c = \partial_n \sum_{i\in c} n_i = \sum_{i\in c} \partial_n n_i + \partial_n f_c = \sum_{i\in c} p_O^i + \chi_c p_N \ .
\end{aligned} \tag{S1}$$

A. Models with correlations

We now deal with the scaling of the basic observables in the model taking into account the correlation between categories growth (model I of the main text).

The correlation appears in the growth of the domain families of different categories. Thus the probability p_O^i that a domain is added to a given family i can be written as

$$p_O^i = \frac{\sum_{j=1}^{f} a_{i,j} n_j - \alpha}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta} \ . \tag{S2}$$

The coordinated growth of functional categories is encoded by the coefficients $a_{i,j}$, responsible for the correlated expansion of evolutionary families i and j (See Equation 1 of the main text). The standard Chinese Restaurant Process (CRP) is obtained by imposing $a_{i,j} = \delta_{i,j}$ (where $\delta_{i,j}$ is equal to 1 if $i = j$ and 0 otherwise). We assume that these coefficients depend only on the functional categories c and c' to which the families i and j belong. The probability of introducing a new domain is given by

$$p_N = \frac{\alpha f + \theta}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta} \ . \tag{S3}$$

1. Model Ia.

We consider a model inspired by ref. [15] (the toolbox model, in which the growth of the number of transcription factors is coupled to the number of added metabolic enzymes), extended to describe a joint partitioning in functional and evolutionary categories. In the original version of the model the average increment of the main observables at each time step is

$$\begin{cases} \Delta n_{met} = \dfrac{U}{n_{met}} \\ \Delta n_{TF} = 1 \ , \end{cases} \tag{S4}$$

and thus $\Delta n_{TF}/\Delta n_{met} = n_{met}/U$, which gives a quadratic scaling for n_{TF} with n_{met}.

Model Ia is an extension of the toolbox model is formulated following equation S2, by using a proper definition of $a_{i,j}$, such as the same equation of the toolbox model is valid. We observe that, for our purpose, the time step of equation we can be defined arbitrarily, as genome growth is eventually parameterized by n. Rewriting the equations as

$$\begin{cases} \Delta n_{met} = n_{met} \\ \Delta n_{TF} = n_{met} \frac{n_{met}}{U} \ , \end{cases} \tag{S5}$$

gives the summed probabilities p_O^i relative to the two categories

$$\begin{cases} p_O^{met} := \displaystyle\sum_{i \in met} p_O^i = \dfrac{n_{met} - \alpha f_{met}}{C(n)} \\ p_O^{TF} := \sum_{i \in TF} p_O^i = \frac{\frac{n_{met}}{U} n_{met} - \alpha f_{TF}}{C(n)} \ , \end{cases} \tag{S6}$$

while

$$p_N = \frac{\alpha f + \theta}{C(n)} \ . \tag{S7}$$

Accordingly, we extend the model to an arbitrary number of families by the choice $a_{i,j} = \frac{n_{met}}{U} \frac{n_i}{n_{TF}}$ if i is a TF family and j a metabolic family and zero otherwise. This gives

$$\begin{cases} p_O^i = \dfrac{\sum_{j \in met} \frac{n_{met}}{U} \frac{n_i}{n_{TF}} n_j - \alpha}{\sum_{i,j=1}^f a_{i,j} n_j + \theta} & \text{if } i \in TF \\ p_O^i = \dfrac{n_i - \alpha}{\sum_{i,j=1}^f a_{i,j} n_j + \theta} & \text{if } i \in met \ . \end{cases} \tag{S8}$$

This model gives the asymptotic quadratic scaling of n_{TF} with n_{met} by definition, using the exact same argument as the toolbox model. Other results have been obtained numerically (see Supplementary Figure S4).

2. *Model Ib.*

This second formulation of a model with correlated recipe (model Ib) imposes a different correlation rule. For example, consider the model involving only two functional categories, transcription factors controlling metabolic processes and metabolic enzymes.

In this variant the coefficients $a_{i,j}$ have both a diagonal and a non diagonal part, $a_{i,j} = \delta_{i,j} + b_{i,j}$. If $b = 0$ the model is the standard Chinese Restaurant Process. For this reason, model Ib is simpler to treat analytically, exploiting previous results. This work focuses mainly on the case $b_{i,j} = n_i/n_{met}$ if i is a family from the functional category of transcription factors and j is a family from the metabolic functional category (and $b_{i,j} = 0$ otherwise).

In this case, the summed probabilities p^i_O relative to the two categories are

$$\begin{cases} p^i_O = \dfrac{n_i + \sum_{j \in met} \frac{n_i}{n_{met}} - \alpha}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta} & \text{if } i \in TF \\ p^i_O = \dfrac{n_i - \alpha}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta} & \text{if } i \in met. \end{cases} \tag{S9}$$

Using the definitions given in Equation S1, one can see that,

$$C(n)\partial_n n_{TF} = n_{TF} + n_{TF} - \alpha f_{TF} + C(n)\partial_n f_{TF} = 2n_{TF} - \alpha f_{TF} + \alpha f_{TF} + \theta \chi_{TF} = 2n_{TF} + \theta \chi_{TF} \ , \tag{S10}$$

while

$$C(n)\partial_n n_{met} = n_{met} - \alpha f_{met} + C(n)\partial_n f_{met} = n_{met} + \theta \chi_{met} \ . \tag{S11}$$

Hence, for large n, since $\partial_n f_c = \chi_c p_N \simeq \alpha f_c$, the terms in the r.h.s. of Equations (S10) and (S11) cancel, giving the effective equation,

$$\frac{dn_{TF}}{dn_{met}} \simeq \frac{2n_{TF}}{n_{met}} \ , \tag{S12}$$

and thus the scaling $n_{TF} \sim n^2_{met}$.

B. Model II (model with evolutionary potentials)

This section presents in more detail the uncorrelated version of the model for the joint scaling (model II), assigning evolutionary potentials [3] ρ_c to the functional categories, related to the probability that a gene added in a functional category is fixed by natural selection. This model is an example of an "absolute recipe", since each category grows with an intrisic rate ρ_c, summing up the growth of the families belonging to the given category. The rate ρ_c acts on family growth through the class-expansion move. The probability of class expansion of a family belonging to the category c is equal to

$$p^i_O = \frac{\rho_{c(i)} n_i - \alpha}{\sum_{j=1}^{f} \rho_{c(j)} n_j + \theta}, \tag{S13}$$

where $\rho_{c(i)} = \rho_c$ if the evolutionary family i belongs to the functional category c. This model assumes that the value of $\rho_c(i)$ depends only on the category to which family i belongs. The probability that a domain belonging to category c is added by class expansion is then

$$p_O^c := \sum_{i \in c} p_O^i = \frac{\rho_c n_c - \alpha f_c}{\sum_{j=1}^f \rho_{c(j)} n_j + \theta}. \tag{S14}$$

Equally, the probability that the new domain is introduced by an innovation move (i.e. it belongs to a new family) is equal to

$$p_N = \frac{\alpha f + \theta}{\sum_{j=1}^f \rho_{c(j)} n_j + \theta}. \tag{S15}$$

Under the assumption (confirmed by empirical data, see main text) that the growth of old functional categories by adding new homology families through the innovation move is uniform (i.e. that $f_c = A_c + \chi_c f$), the probability that a new family belonging to the category c is added by an innovation move is

$$p_N^c := \chi_c p_N = \chi_c \frac{\alpha f + \theta}{\sum_{j=1}^f \rho_{c(j)} n_j + \theta} = \frac{\alpha f_c + \theta \chi_c}{\sum_{j=1}^f \rho_{c(j)} n_j + \theta}. \tag{S16}$$

Evolutionary potentials can reproduce the combined scaling laws at finite sizes.

We tested this model by a combination of mean-field analytical arguments and direct simulation. The mean-field equations are obtained from Equation S1 by using Equations S13 and S15. The equation for the growth of the mean number of members n_c of a functional category can be obtained simply by summing on the homology families that belong to a given category,

$$\partial_n n_c = \frac{\rho_c n_c + \theta \chi_c}{C(n)} \ , \tag{S17}$$

where $C(n) \simeq \sum_i \rho_i n_i$. If $C(n) \sim n$, equation (S17) corresponds to the evolution equation written by Molina and Nimwegen. Simulations of this model (see Supplementary Figure S7) confirm that this is the case. Thus, the mean-field argument predicts that this model can reproduce both scaling laws.

Also note that a rescaling of $C(n)$ is equivalent to a rescaling of α. Indeed, for large n, $p_N \simeq \alpha f / C(n)$ (and $p_O = 1 - p_N$), so imposing $C(n) \simeq qn$ is equivalent to dividing α by the constant factor q. Thus, one can choose $q = 1$ without loss of generality (by a rescaling of all the ρ_c), and the solution for the population of a functional category will be $n_c \sim n^{\rho_c/q}$ as in the Molina/Nimwegen model, and thus $\zeta_c = \rho_c/q$

On the other hand, an important point regarding this model is that, asymptotically for any choice over the ρ_c set, the maximum large-n exponent observed will be 1, Indeed, we can use the approximation $C(n) = \sum_i \rho_i N_i \sim \rho_{\max} n^{\rho_c/q}$, but $C = qn$, so that $q = \rho_{c_{max}}$. This means that an exponent close to 2, such as that observed for transcription factors can only be obtained in a transient regime of the model. Furthermore, the change of the evolutionary potential of one functional category has repercussions on the other categories, as it implies a change in the normalization costant C. These facts make a direct identification of the value of the evolutionary potential with an intrinsic propery of a single functional category difficult. They also make the direct identification of evolutionary potentials less straightforward (as it requires an arbitrary rescaling).

However, the above remarks have little practical importance, and the large-n behaviour of the model does not really affect its performance at the relevant values of n. Numerical simulations show that at the empirical genome sizes, the scaling behaviour of the model can reproduce rather well the empirical one. For simplicity we have restricted to three main categories (transcription factors, metabolic genes and "others") and we verified that in practice it is not hard to find a parameter set in good agreement with the empirical data on protein domains (Supplementary Figure S3). The general number of parameters to adjust increases with the number of functional categories that one needs to consider.

S2. EXPONENTS OF FAMILY SIZE DISTRIBUTION HISTOGRAMS

This section discusses the family size distribution histograms, as obtained from the mean-field approach. To fix the ideas, we will focus on model Ib, where the mean-field equations can exploit the known results from the CRP. It is possible to write a mean-field "flux equation" for the histograms [14], which implements the fact that each duplication adds a family with one extra member to the histogram count and subtracts a family with its previous population,

$$\partial_n f(d,n) = p_O(d-1,n) f(d-1,n) - p_O(d,n) f(d,n) + p_N \delta_{d,1} \qquad \text{(S18)}$$

where $p_O(d,n) = \frac{d-\alpha}{n+\theta}$ is the probability that a family with d domains add a new duplicated member. The term $p_N = \frac{\alpha f + \theta}{n+\theta}$ contains the innovation probability contributing to the growth of the number of families with one member. Note that the flow between families can be written as

$$\sum_{i \in \left\{ \substack{\text{families with} \\ j \text{ domains}} \right\}} \partial_n n_i = (d-\alpha) \frac{f(d,n)}{n+\theta}.$$

This equation requires an assumption on $f(d,n)$ in order to be solved. We assume the ansatz $f(d,n) = P(d)f(n)$ which is justified by both simulation and empirical data [14]. Using the fact that $\partial_n f(n) = p_N$, combined with Equation S18 gives the following equation for the probability of a family to have d members

$$\alpha P(d) = (d - 1 - \alpha)P(d-1) - (d-\alpha)P(d) \quad , \tag{S19}$$

which can be solved in discrete or continuous d to get

$$P(d) \sim \left(\frac{1}{d}\right)^{1+\alpha} \quad . \tag{S20}$$

This predicts the asymptotic behaviour of data and simulations (see Figure 6) with $\beta = \alpha$, where β is the asymptotic exponent of the family size distribution.

Let us now turn to the same distribution, restricted to transcription factors. In model Ib, the flux from transcription factor families caused by family expansion is caused by two separate contribution, the CRP standard one, plus additions of transcription factors to an existing family caused by the addition of a metabolic enzyme

$$p_O^i(n) = \frac{1}{C(n)}\left[(n_i - \alpha) + \frac{n_i}{n_{met}} n_{met}\right], \text{ if } i \in TF \tag{S21}$$

i.e.

$$p_O^i(n) = \frac{1}{C(n)}\left[2n_i - \alpha\right], \text{ if } i \in TF. \tag{S22}$$

Thus, for the transcription factor families, the probability that a domain is added to a family with d members will be

$$p_O^{TF}(d,n) = \frac{1}{C(n)}\left[2d - \alpha\right] \quad . \tag{S23}$$

The quantity $p_O^{TF}(d,n)$ is the probability that a new transcription factor domain is added to a family with d members. The flux equation for TF families can be obtained by substituting equation S23 in equation S18, (for $d > 1$)

$$C(n)\partial_n f_{TF}(d,n) = \left[2(d-1) - \alpha\right] f_{TF}(d-1,n) - \left[2d - \alpha\right] f_{TF}(d,n) \tag{S24}$$

This is solved using the usual ansatz $f_{TF}(d,n) = P_{TF}(d)f_{TF}(n)$ (as explained above it is confirmed by both data and simulations). Using $f_{TF}(n) = \chi_{TF} f(n)$, leads to the equation

$$\alpha P_{TF}(d) = (2d - 2 - \alpha)P_{TF}(d) - (2d - \alpha)P_{TF}(d) \quad , \tag{S25}$$

which gives:

$$P(d)_{TF} \sim \left(\frac{1}{d}\right)^{1+\frac{\alpha}{2}} \quad , \tag{S26}$$

that is $\beta_{TF} = \alpha/2 = \beta/2$. In the above calculation we have supposed again that the number of transcription factors is small with respect to to the total number of metabolic enzymes.

Furthermore, it can be argued that this fact is more general. Indeed, each time the per-family duplication probability for the TF functional category will have the form

$$p_O^i \simeq 2n_i \ ,$$

when family i belongs to TF category, the coefficient 2 will appear in the equation for $P(d)_{TF}$ modifying the exponent. In particular, this will also be true for models Ia (generalizing the toolbox model) and II (generalizing evolutionary potentials).

In other words, each time a functional category scales with a given exponent, it can be argued on rather general grounds that the exponent of the population histograms of the homology families that form it will be affected. It is possible to to generalize this argument, and find a precise relationship between the scaling exponent of a category and the family population histogram (restricted to the same category). In other words, if ζ_c is the scaling exponent of the category c and β_c is the exponent of the cumulative distribution histogram for the families belonging to category c, that is (see Equation S26):

$$P(d)_c \sim \left(\frac{1}{d}\right)^{1+\beta_c} \quad ,$$

we suggest that $\beta_c = \beta/\zeta_c$. We tested this prediction in empirical data plotting $1/\beta_c$ versus ζ_c in Figure 6 (Pearson correlation coefficient 0.47).

S3. COMPARISON OF MODELS BY NUMERICAL SIMULATION

A. Correlated and absolute recipes

This section compares the correlated duplication and the evolutionary potential model variants. We considered a three categories model (TF, Metabolic and "other").

The evolutionary potential model needs to supply three parameters ρ_c, while the correlated model needs to supply the correlation law between categories (a_{ij}). We impose a correlation only between transcription factor and metabolic families with the correlated model Ib prescription, i.e.

$$a_{ij} = n_i/n_{met}, \tag{S27}$$

where i is a TF family and j Metabolic, $a_{ij} = 0$ (no correlation) otherwise.

Figure S3 summarizes the results of this comparison. The correlated duplication model performs better in reproducing the behavior of the transcription factor category (both scaling law and histograms). Both models are unsatisfactory in reproducing the family population histogram of the metabolism families. This is probably caused by the fact that neither model include a correlation between metabolic families (Figure 7).

Figure S7 illustrates the behaviour of the normalization function $C(n)$. $C(n)$ is linear with n in the range of empirical genome sizes (although the slope is not exactly 1). It becomes nonlinear at larger sizes, and its linear behavior is restored only at very large values of n.

B. Model I can reproduce a set of different exponents

Extending a model (with absolute or correlated recipe) to a large number of categories is not a simple task. In the case of an absolute recipe model, adding a new category c' (and thus introducing a new evolutionary potential $\rho_{c'}$) generally requires, in order preserve the scaling of all the categories, a tuning of all the evolutionary potentials (both the old ones and the new one). This is due to the fact that all the evolutionary potentials appear in the normalization constant $C(n)$ in the growth equation of each category (Equation (S13). In a model with a correlated recipe, the main problem is related to the fact that the interaction laws between categories are not known, they can be complex and possibly include feedback.

In order to produce the proof of principle that a model with correlated recipes can work with more than three categories, we considered a trivial generalization of model Ib to multiple categories that are slaved to a main one, and considered the question of whether this model would be able to reproduce an arbitrary set of scaling exponents for the categories.

We consider a correlation matrix $a_{i,j}$ of the form $\delta_{i,j} + b_{i,j}$, where $b_{i,i} = 0$. This model deals with $\mathcal{C}$+1 categories, the *met* category (in analogy with model Ib defined in the main text, this is a category whose growth is not conditioned to the others), and an additional set of $\mathcal{C}$ categories labeled from 1 to $\mathcal{C}$. The non diagonal correlation coefficients $b_{i,j}$ are zero if family i belongs to the *met* category, and $\gamma_{c(i)} n_i / n_{met}$ if family i belongs to category c, different from *met*, and j belongs to the *met* category. Substituting this choice in equation S9, gives

$$\frac{dn_c}{dn_{met}} = (1 + \gamma_c) \frac{n_c}{n_{met}} \qquad \text{(S28)}$$

and thus

$$n_c \sim n_{met}^{1+\gamma_c} \ . \tag{S29}$$

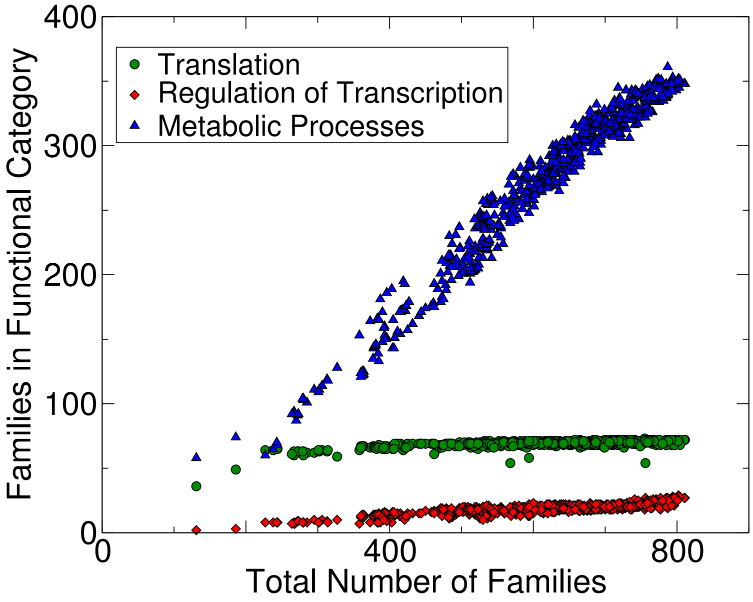
Supplementary Figure S8 shows simulations from a model with $10 + 1$ categories. The model is able to reproduce an arbitrary set of exponents. We observe that this version has similar problems as the model with evolutionary potentials, as, in absence of a biological underlying model, it needs the tuning of a set of parameters to reproduce the scaling laws. The fitted exponent is typically different from $1 + \gamma_c$, specifically it seems to be closer to one. We interpret this as a finite size effect, due to the fact that the contribution of innovation to the scaling exponents is relevant.

S4. DETAILS OF TF-DOMAIN SUPERFAMILY SCALING

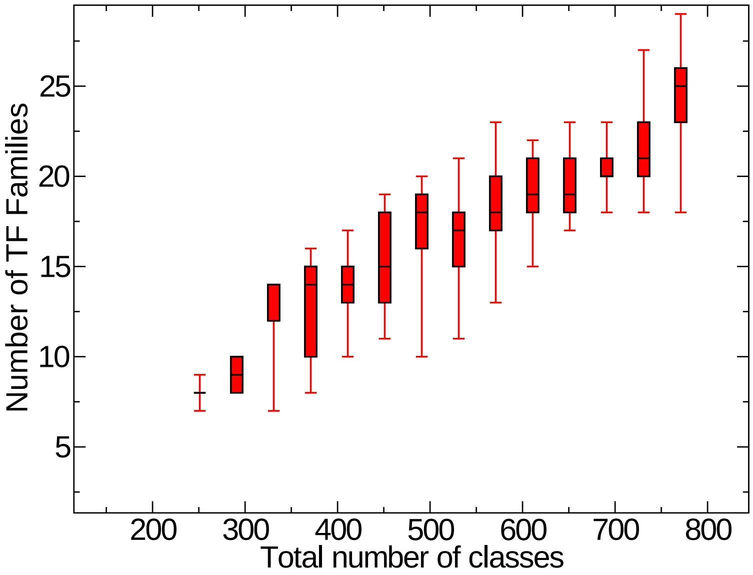
We observe that the quadratic (or very nearly so) scaling for transcription factors is clearly visible at in the two most populated families of transcription factor DNA-binding domains (Homeodomain-like and Winged-helix), which have a rather clean slope (see Supplementary Figure S10). In fact, three families present a clearly observable scaling alone (Homeodomain-like, Winged-helix and C-terminal), but just the first two follow a very nearly quadratic scaling.

Note however that removing the six most populated TF families, representing 80% of the total TF-domain population, the remaining ones considered together still present a scaling when added up, but with exponent $\simeq 0.9$ (see Supplementary Figure S11). This indicates that the collective scaling of TF families cannot be entirely recunducted to properties of the most populated ones, but these are the families responsible for the *quadratic* scaling.

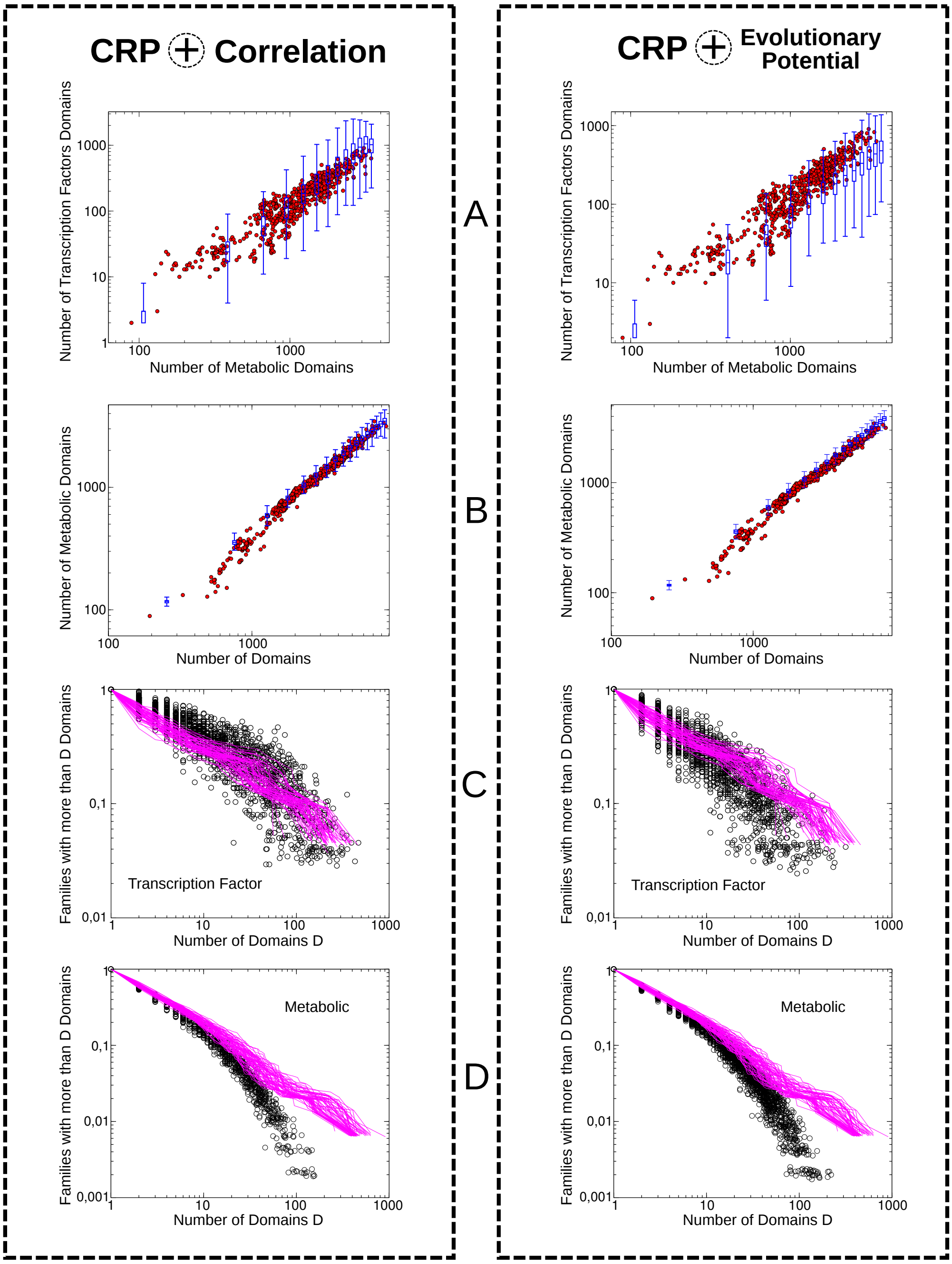
Thus, the "pure" quadratic scaling is observable in the largest transcription factor families. Collecting all the families, wemeasure a lower exponent in empirical data (close to 1.6). Supplementary Figure S11 explains this behavior, showing the total contribution of the smaller transcription factor families. These families collectively show a lower exponent (close to 1). Thus, we can interpret the lower collective exponent as an effect of family size (i.e., in the language of statistical mechanics, a "finite-size" effect) connected to the fact that for smaller family size, the innovation move is more relevant and thus the family expansion process is slower. The same effect is present in our simulations (see Supplementary Figure S12.)



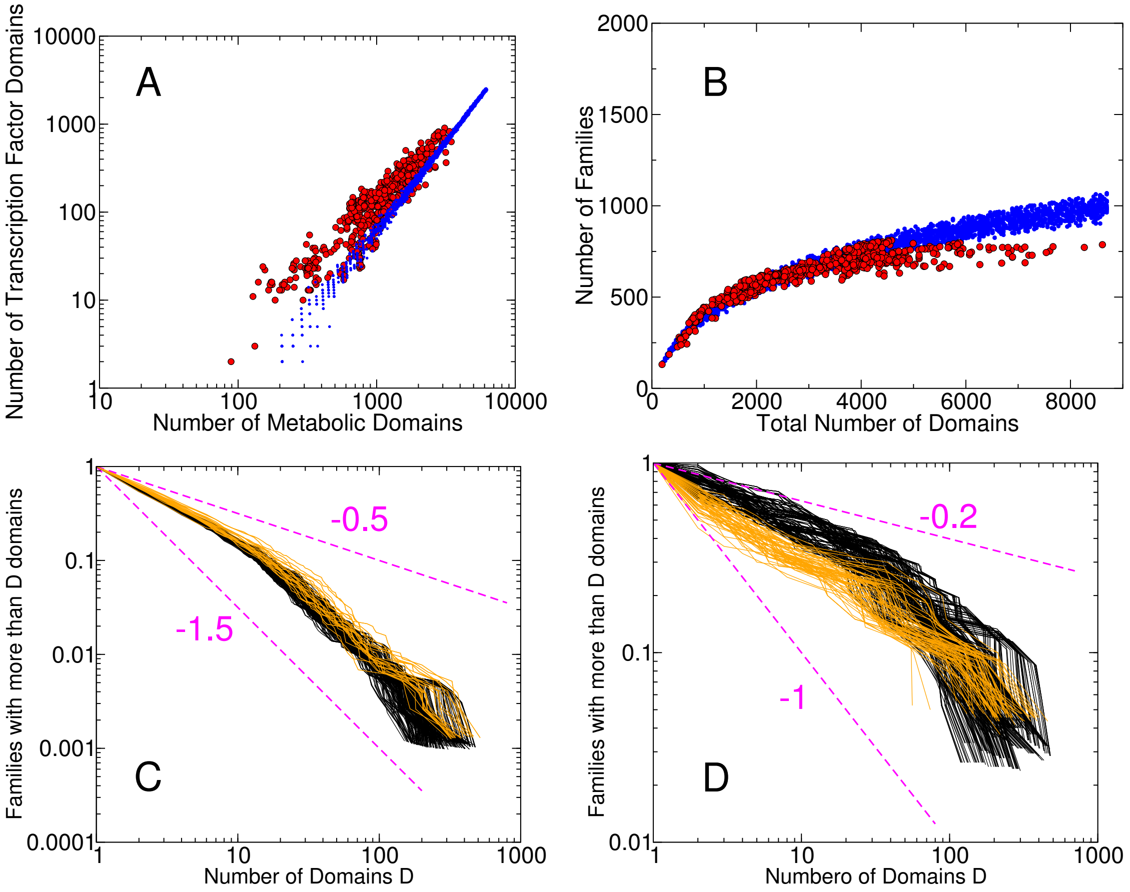
Supplementary Figure S1: **Scaling of the number of families in the three main functional categories.** Linear scaling behaviour of the number of families in three important functional categories versus total number of families from empirical data (for 753 bacteria in the SUPERFAMILY database). The slopes for the three linear laws are 0.01 (Translation), 0.03 (Regulation of Transcription) and 0.47 (Metabolic Processes).



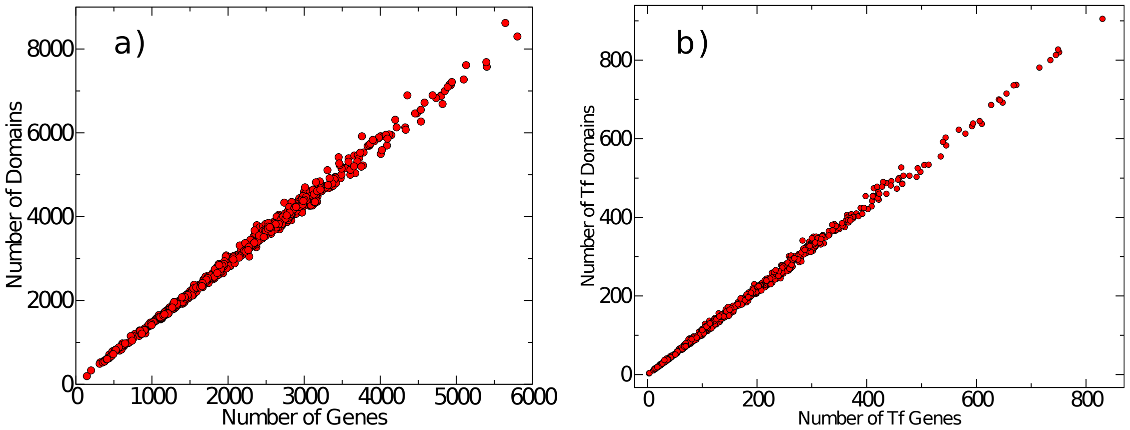
Supplementary Figure S2: **Transcription factor families.** Boxplot of the number of transcription factor domain families versus total number of domain families (data from 753 SUPERFAMILY bacteria). There appears to be a roughly linear scaling. This means that the number of TF domain families is compatible with a null hypothesis of independent addition model. Charoensawan *et al* [22] propose that the number of TF families follows a linear scaling with genome *size*. If this were to be the case, the innovation dynamics of transcription factor families should be distinct form other families. In fact, if $f_{TF}(n) \sim n$, since the total number of families is sublinear, $f(n) \sim n^{\alpha}$ in the CRP (Figure 1), then one would have $f_{TF}\ f^{2-\alpha}$, which is not confirmed by the SUPERFAMILY data analyzed here.



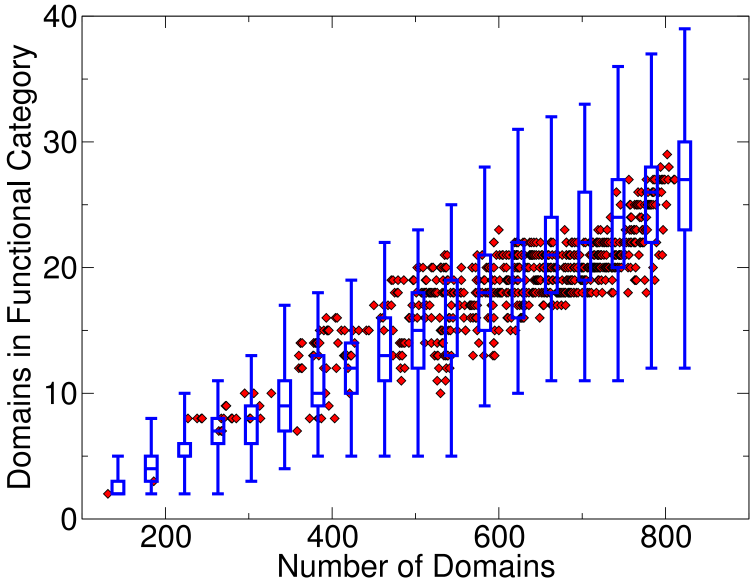
Supplementary Figure S3: **Comparison between models Ib and II.** Comparison between simulation of the correlated duplication model Ib (left panel) and evolutionary potentials (right panel) model variants with empirical data. Simulations are run at $\alpha = 0.3$ and $\theta = 140$. (a) Number of TFs domains vs. number of metabolic domains (the blue boxplot corresponds to simulations, red circles to empirical data). (b) Number of metabolic domains vs. total number of domains (the blue boxplot corresponds to simulations, red circles to empirical data). (c) Family population histograms restricted to the transcription factor functional category (black circles are simulations, magenta lines empirical data). (d) Family population histograms restricted to the metabolism functional category (black circles are simulations, magenta lines empirical data).



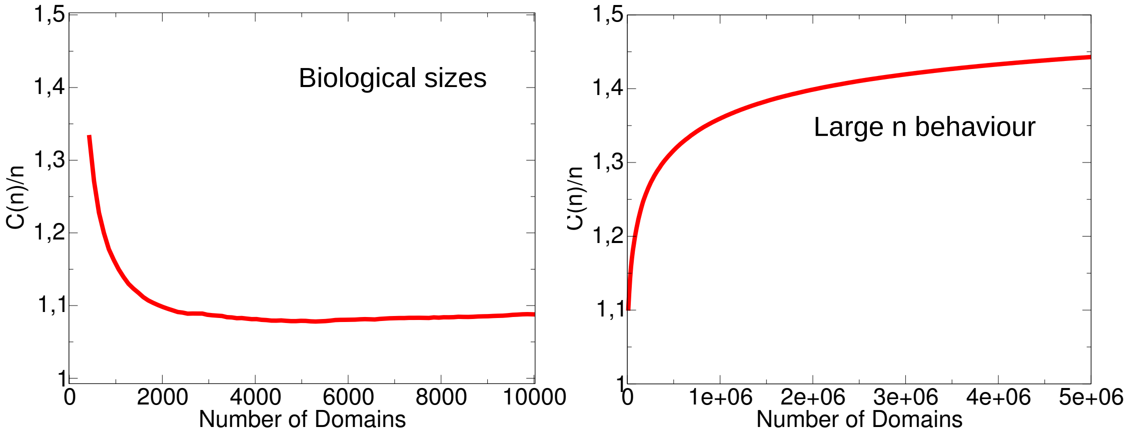
Supplementary Figure S4: **Simulations of the correlated duplication model Ia for two categories (transcription factors and metabolic enzymes).** The plots are obtained from 1000 realizations with $\alpha = 0.3$, $\theta = 140$ and $U = 7000$. The observables are the same as in figure S3. (a) scaling of the number of transcription factors with the number of metabolic enzymes. (b) Number of families as a function of genome size n. (c) Family population (cumulative) histograms. (c) Family population histograms restricted to the families belonging to the transcription factor functional category.



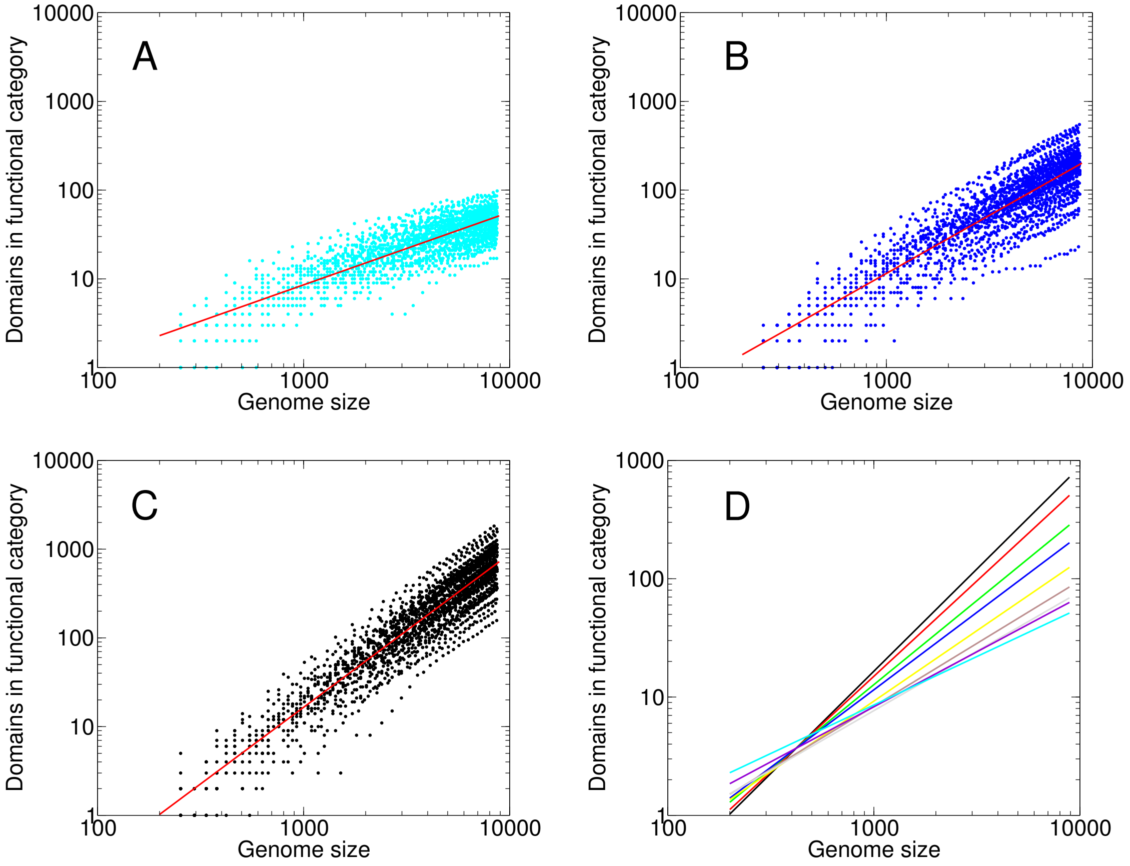
Supplementary Figure S5: **Linear relation between the number of domains and the number of genes.** (a) Number of Domains vs. number of protein-coding genes for the 753 bacteria in the SUPERFAMILY database. There are, on average, 1.45 domains per gene. (b) Linear scaling behaviour of the number of TF domains vs. number of TF genes. There are, on average, 1.09 TF domains in a TF gene.



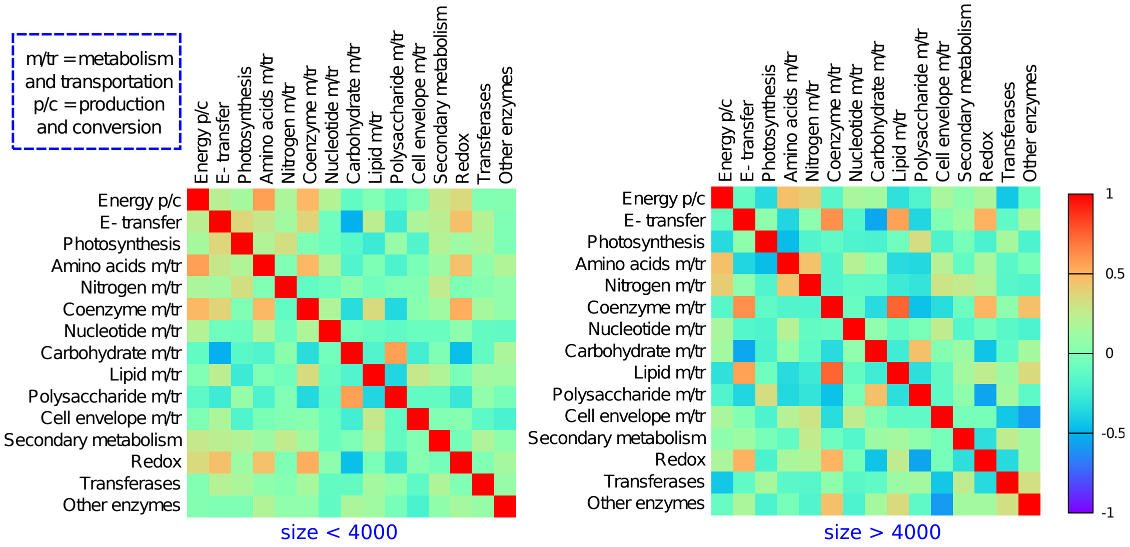
Supplementary Figure S6: **Simulation of the number of transcription factor families.** Comparison between empirical data and simulations of the number of transcription factor domain families plotted against total number of families. The scaling is empirically linear, i.e. the number of TF domain families is reproduced by a null hypothesis of independent addition model. The choice of the parameter is 0.035.



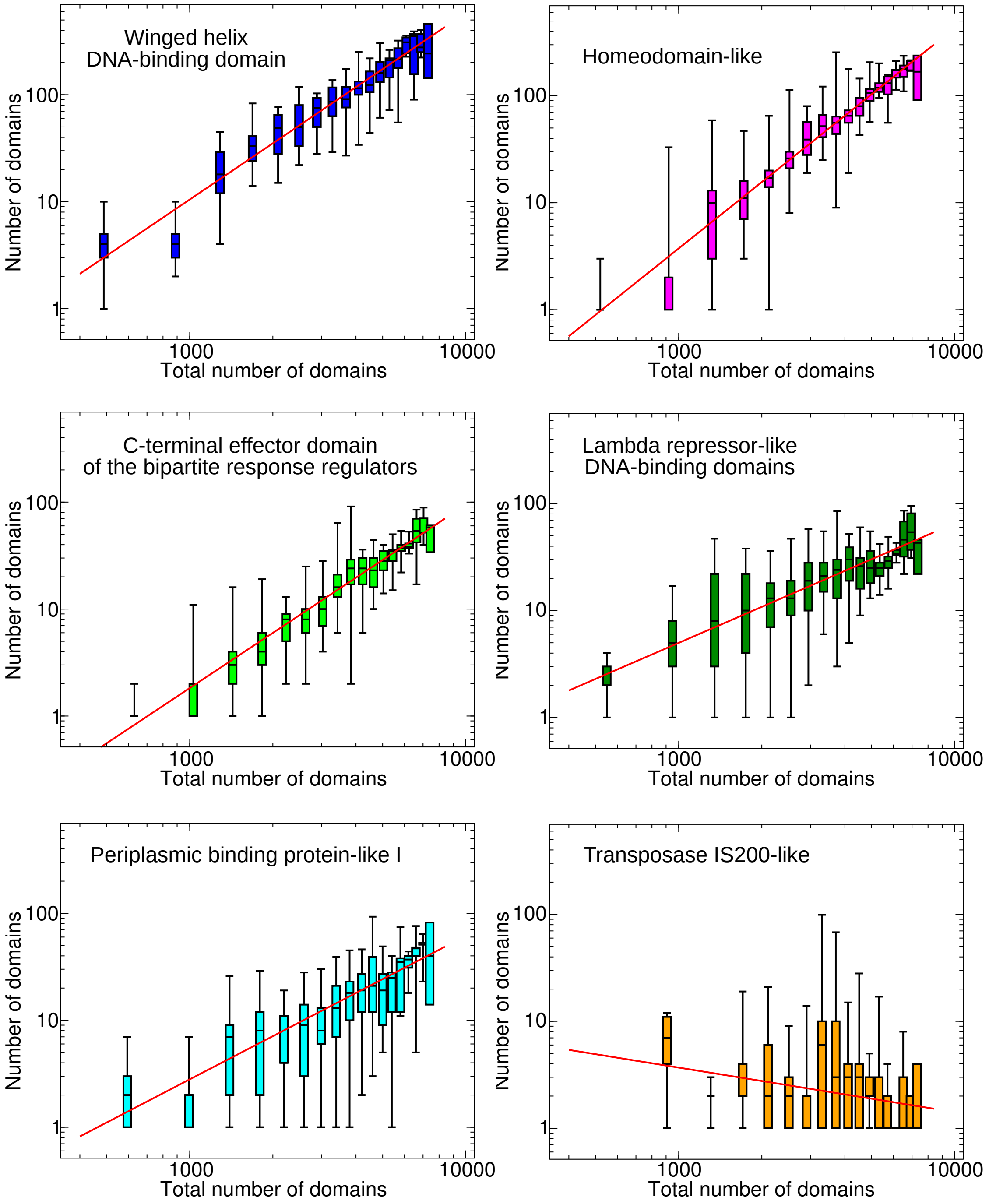
Supplementary Figure S7: **Normalization constant inthe model with evolutionary potentials (model II).** Behavior of the ratio $C(n)/n$, where $C(n)$ is the normalization factor for the evolutionary potential model. Data from simulations with three categories run at parameters $\alpha = 0.3$ and $\theta = 140$. $C(n)$ is linear with n in the range of empirical genome sizes, it then looses linearity, to become linear only asymptotically.



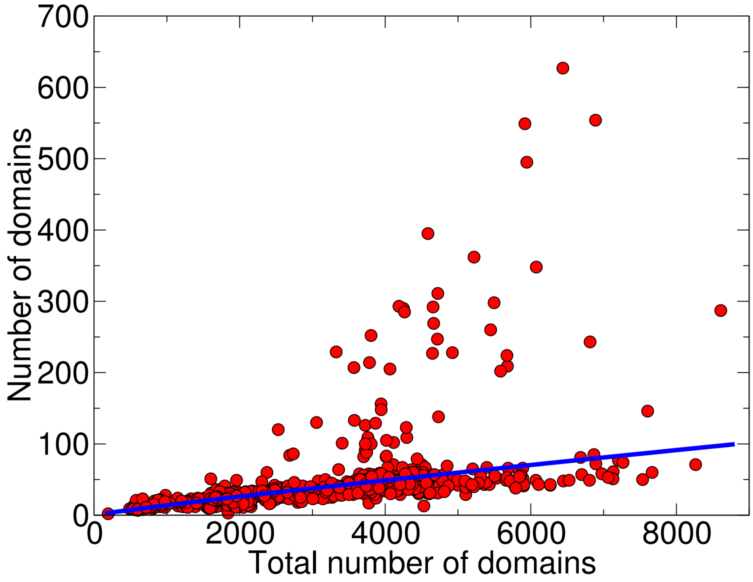
Supplementary Figure S8: **Simulation of model Ib with** $10+1$ **categories.** 10 categories are slaved to one master category with different correlation laws, which determine the observed exponents). Panel A, B and C show the simulations of the population of three categories (respectively with γ_c equal to 1, 0 and -0.7). The red lines are power-law fits of the simulated data. Panel D shows the power-law fits of the simulated data for all ten categories.



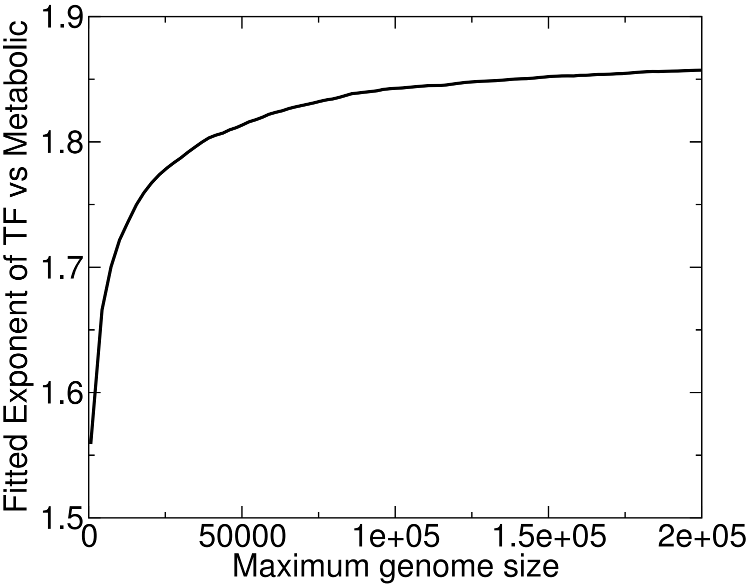
Supplementary Figure S9: **Correlation matrix for two sets of genomes with different sizes.** Left panel: Correlation matrix for genomes with size $<$ 4000. Right panel: Correlation matrix for genome with size $>$ 4000. The correlations do not depend on size.



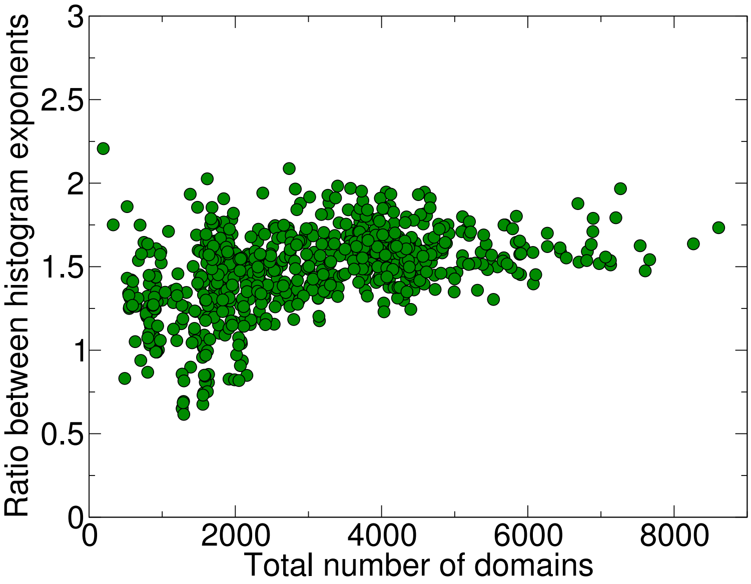
Supplementary Figure S10: **Most populated transcription factor superfamilies.** Boxplots for the population of the six most populated superfamilies of TF DNA-binding domains (y-axis in each panel) versus number of domains of each genome (x-axis in each panel). The presence of scaling laws appears likely for the three most populated families and arguable for the first five. Red lines represent best power law fit (1.8 for Winged Helix ,2.1 for Homeodomain-like and 1.7 for C-terminal effector)



Supplementary Figure S11: **Scaling of the least populated transcription factor superfamilies.** Collective scaling of the number of transcription factor domains after removing the six globally most populated families. While a few genomes show large fluctuations from the typical trend, a clear scaling is still observable for most genomes, with a fitted exponent equal to 0.9



Supplementary Figure S12: MARCO **Finite-size effects on the scaling exponent ζ_{TF} for transcription factors in simulations of model Ib.** The plot shows the fitted exponent (y-axis) from the curve of the number of transcription factor domains versus the number of metabolic enzymes in 500 simulated realizations of model Ib with parameter $\alpha = 0.3$ and $\theta = 140$. Each point on the x-axis corresponds to simulated data stopped at a given size n. The mean-field prediction ($\zeta_{TF} = 2$) is reached only in the limit $n \to \infty$. This plot shows that the fitted exponent 1.6 (instead of 2) for the growth of transcription factors vs metabolic domains is due to a finite-size effect of a process that produces an exponent 2 in the large-n limit. The same effect is present in models Ia and II.



Supplementary Figure S13: **Ratio between exponents of family population histograms.** The plot reports the ratio β/β_{TF} between the exponent of the total family population histograms and the histograms restricted to the transcription factor families (see Figure 5 in the main text), as a function of genome size. The values of the ratio are distributed around 1.6 and the fluctuation range decreases with increasing genome size.

Supplementary Table S1: **Fitted values of χ_c and offsets A_c from f_c vs f for the ten largest functional categories**

	A_c	χ_c	Reduced chi square
Transcription Factors	2.2 ± 0.4	0.0267 ± 0.0006	4.5
Translation	61.0 ± 0.35	0.0133 ± 0.0006	3.9
Small molecule binding	3.0 ± 0.2	0.01 ± 0.0002	0.9
Nucleotide transport and metabolism	5.6 ± 0.3	0.02 ± 0.0005	3.1
DNA replication/repair	9.5 ± 0.6	0.0437 ± 0.0009	9.8
Inorganic ion transport and metabolism	0.2 ± 0.4	0.0272 ± 0.0005	3.5
Redox	-7.6 ± 0.5	0.0592 ± 0.0008	7.9
Transferases	5.3 ± 0.2	0.0213 ± 0.0004	1.6
Other enzymes	-14.8 ± 1.1	0.155 ± 0.002	35.7
Signal transduction	-3.2 ± 0.3	0.0282 ± 0.0005	3.3

The number of evolutionary families belonging to a functional category follows a linear law in empirical data. The table reports fits of $f_c = A_c + \chi_c f$ from the plots in Figure 2 of the main text, where f_c represents the number of families in category c on all genomes and f is the total number of families on the genome. The third column is the reduced chi square.

Supplementary Table S2: **Data of fitted exponents from Figure 6 of the main text, for the ten largest functional categories**

	ζ_c	β_c
Transcription Factors	1.6 ± 0.02	0.47 ± 0.01
Translation	0.176 ± 0.003	1.46 ± 0.02
Small molecule binding	0.918 ± 0.006	0.25 ± 0.01
Nucleotide transport and metabolism	0.61 ± 0.01	0.71 ± 0.01
DNA replication/repair	0.54 ± 0.01	0.9 ± 0.01
Inorganic ion transport and metabolism	1.40 ± 0.02	0.46 ± 0.01
Redox	1.3 ± 0.01	0.52 ± 0.02
Transferases	1.09 ± 0.01	0.43 ± 0.01
Other enzymes	1.09 ± 0.01	0.64 ± 0.01
Signal transduction	1.77 ± 0.03	0.4 ± 0.01

Supplementary Table S3: **Correlation coefficients between the populations of metabolic functional categories**

	En	e-	Ph	Aa	N	Co	Nu	Ca	Li	Ps	Ce	2M	Rx	Tr	Ot
En	1	0.14	0.07	0.55	0.23	0.36	0.19	−0.06	−0.08	−0.14	0.02	0.22	0.31	−0.10	−0.004
e-	0.14	1	0.29	0.15	0.11	0.43	−0.09	−0.52	0.35	−0.29	0.13	0.19	0.47	0.09	0.05
Ph	0.07	0.29	1	0.12	0.21	−0.02	−0.09	−0.16	−0.21	0.14	−0.18	0.15	0.06	0.16	−0.05
Aa	0.55	0.15	0.12	1	0.08	0.39	0.19	−0.14	−0.07	−0.22	0.01	0.07	0.40	0.02	0.14
N	0.23	0.11	0.21	0.08	1	−0.13	−0.08	−0.003	−0.14	−0.09	0.09	0.26	0.04	−0.03	−0.02
Co	0.36	0.43	−0.02	0.39	−0.13	1	0.14	−0.33	0.44	−0.37	−0.04	0.08	0.51	0.12	0.16
Nu	0.19	−0.09	−0.09	0.19	−0.08	0.14	1	−0.03	−0.09	−0.10	−0.02	−0.10	0.03	−0.11	−0.13
Ca	−0.06	−0.52	−0.16	−0.14	−0.003	−0.33	−0.03	1	−0.20	0.53	−0.18	0.02	−0.46	−0.11	0.16
Li	−0.08	0.35	−0.21	−0.07	−0.14	0.44	−0.09	−0.20	1	−0.35	0.15	0.18	0.06	0.13	0.20
Ps	−0.14	−0.29	0.14	−0.22	−0.09	−0.37	−0.10	0.53	−0.35	1	−0.12	−0.05	−0.36	0.09	−0.07
Ce	0.02	0.13	−0.18	0.01	0.09	−0.04	−0.02	−0.18	0.15	−0.12	1	−0.0002	0.01	−0.22	−0.31
2M	0.22	0.19	0.15	0.07	0.26	0.08	−0.10	0.02	0.18	−0.05	−0.0002	1	−0.11	0.20	0.08
Rx	0.31	0.47	0.06	0.40	0.04	0.51	0.03	−0.46	0.06	−0.36	0.01	−0.11	1	−0.10	0.14
Tr	−0.10	0.09	0.16	0.02	−0.03	0.12	−0.11	−0.11	0.13	0.09	−0.22	0.20	−0.10	1	0.17
Ot	−0.004	0.05	−0.05	0.14	−0.02	0.16	−0.13	0.16	0.20	−0.07	−0.31	0.08	0.14	0.17	1

Pearson's correlation coefficients between the populations of 24 different metabolic functional categories from the SUPERFAMILY database for 753 bacteria. Correlations are calculated from fluctuations of categories from the average trend (see Methods). Both correlation and anticorrelation are present between categories. Metabolism categories are highly (anti-)correlated. We used the following short forms for the metabolic functional categories: En = Energy p/c, e- = Electrons transfer, Ph = Photosynthesis, Aa = Amino acids m/tr, N = Nitrogen m/tr, Co = Coenzyme m/tr, Nu = Nucleotide m/tr, Ca = Carbohydrate m/tr, Li = Lipid m/tr, Ps = Polysaccharide m/tr, Ce = Cell envelope m/tr, 2M = Secondary metabolism, Rx = Redox, Tr = Transferases, Ot = Other enzymes. Where m/tr stands for "metabolism and trasportation" and p/c means "production and conversion".

Supplementary Table S4: **P-Values of correlation coefficients between the populations of metabolic functional categories**

	En	e-	Ph	Aa	N	Co	Nu	Ca	Li	Ps	Ce	2M	Rx	Tr	Ot
En	0	$\mathbf{5 \cdot 10^{-5}}$	**0.02**	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	0.05	**0.01**	$\mathbf{4 \cdot 10^{-5}}$	0.26	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{4 \cdot 10^{-3}}$	0.46
e-	$\mathbf{5 \cdot 10^{-5}}$	0	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-5}}$	$\mathbf{1 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	$\mathbf{8 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{3 \cdot 10^{-4}}$	$\mathbf{1 \cdot 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{7 \cdot 10^{-3}}$	0.08
Ph	**0.02**	$\mathbf{< 10^{-6}}$	0	$\mathbf{1 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	0.29	$\mathbf{2 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-4}}$	$\mathbf{< 10^{-6}}$	$\mathbf{5 \cdot 10^{-5}}$	0.06	$\mathbf{2 \cdot 10^{-5}}$	0.08
Aa	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-5}}$	$\mathbf{1 \cdot 10^{-3}}$	0	**0.02**	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-6}}$	$\mathbf{4 \cdot 10^{-5}}$	**0.02**	$\mathbf{< 10^{-6}}$	0.39	**0.03**	$\mathbf{< 10^{-6}}$	0.28	$\mathbf{5 \cdot 10^{-5}}$
N	$\mathbf{< 10^{-6}}$	$\mathbf{1 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	**0.02**	0	$\mathbf{2 \cdot 10^{-4}}$	**0.01**	0.47	$\mathbf{7 \cdot 10^{-5}}$	$\mathbf{5 \cdot 10^{-3}}$	$\mathbf{8 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	0.13	0.18	0.31
Co	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	0.29	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-4}}$	0	$\mathbf{1 \cdot 10^{-4}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	0.13	**0.02**	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-4}}$	$\mathbf{4 \cdot 10^{-6}}$
Nu	$\mathbf{< 10^{-6}}$	$\mathbf{8 \cdot 10^{-3}}$	$\mathbf{2 \cdot 10^{-3}}$	$\mathbf{2 \cdot 10^{-6}}$	**0.01**	$\mathbf{1 \cdot 10^{-4}}$	0	0.20	$\mathbf{5 \cdot 10^{-3}}$	$\mathbf{3 \cdot 10^{-3}}$	0.26	$\mathbf{3 \cdot 10^{-3}}$	0.17	$\mathbf{8 \cdot 10^{-3}}$	$\mathbf{9 \cdot 10^{-5}}$
Ca	0.05	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{4 \cdot 10^{-5}}$	0.47	$\mathbf{< 10^{-6}}$	0.20	0	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	0.30	$\mathbf{< 10^{-6}}$	$\mathbf{8 \cdot 10^{-4}}$	$\mathbf{7 \cdot 10^{-6}}$
Li	**0.01**	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	**0.02**	$\mathbf{7 \cdot 10^{-5}}$	$\mathbf{< 10^{-6}}$	$\mathbf{5 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	0	$\mathbf{< 10^{-6}}$	$\mathbf{3 \cdot 10^{-5}}$	$\mathbf{< 10^{-6}}$	0.06	$\mathbf{3 \cdot 10^{-4}}$	$\mathbf{2 \cdot 10^{-6}}$
Ps	$\mathbf{4 \cdot 10^{-5}}$	$\mathbf{< 10^{-6}}$	$\mathbf{2 \cdot 10^{-4}}$	$\mathbf{< 10^{-6}}$	$\mathbf{5 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	$\mathbf{3 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	0	$\mathbf{5 \cdot 10^{-4}}$	0.07	$\mathbf{< 10^{-6}}$	$\mathbf{6 \cdot 10^{-3}}$	**0.03**
Ce	0.26	$3 \cdot 10^{-4}$	$< 10^{-6}$	0.39	$8 \cdot 10^{-3}$	0.13	0.26	$< 10^{-6}$	$3 \cdot 10^{-5}$	$5 \cdot 10^{-4}$	0	0.50	0.38	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$
2M	$\mathbf{< 10^{-6}}$	$\mathbf{1 \cdot 10^{-6}}$	$\mathbf{5 \cdot 10^{-5}}$	**0.03**	$\mathbf{< 10^{-6}}$	**0.02**	$\mathbf{3 \cdot 10^{-3}}$	0.30	$\mathbf{< 10^{-6}}$	0.07	0.50	0	$\mathbf{8 \cdot 10^{-4}}$	$\mathbf{< 10^{-6}}$	**0.01**
Rx	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	0.06	$\mathbf{< 10^{-6}}$	0.13	$\mathbf{< 10^{-6}}$	0.17	$\mathbf{< 10^{-6}}$	0.06	$\mathbf{< 10^{-6}}$	0.38	$\mathbf{8 \cdot 10^{-4}}$	0	$\mathbf{3 \cdot 10^{-3}}$	$\mathbf{4 \cdot 10^{-5}}$
Tr	$\mathbf{4 \cdot 10^{-3}}$	$\mathbf{7 \cdot 10^{-3}}$	$\mathbf{2 \cdot 10^{-5}}$	0.28	0.18	$\mathbf{2 \cdot 10^{-4}}$	$\mathbf{8 \cdot 10^{-4}}$	$\mathbf{8 \cdot 10^{-4}}$	$\mathbf{3 \cdot 10^{-4}}$	$\mathbf{6 \cdot 10^{-3}}$	$\mathbf{< 10^{-6}}$	$\mathbf{< 10^{-6}}$	$\mathbf{3 \cdot 10^{-3}}$	0	$\mathbf{1 \cdot 10^{-6}}$
Ot	0.46	0.08	0.08	$\mathbf{5 \cdot 10^{-5}}$	0.31	$\mathbf{4 \cdot 10^{-6}}$	$\mathbf{9 \cdot 10^{-5}}$	$\mathbf{7 \cdot 10^{-6}}$	$\mathbf{2 \cdot 10^{-6}}$	**0.03**	$\mathbf{< 10^{-6}}$	**0.01**	$\mathbf{4 \cdot 10^{-5}}$	$\mathbf{1 \cdot 10^{-6}}$	0

P-values of the Pearson's correlation coefficients between the populations of 24 different metabolic functional categories from the SUPERFAMILY database for 753 bacteria (the most significant values are in boldface). Correlations are calculated from fluctuations of categories from the average trend (see Methods). The (anti-)correlation is statistically significant for the most of the metabolic categories. We used the following short forms for the metabolic functional categories: En = Energy p/c, e- = Electrons transfer, Ph = Photosynthesis, Aa = Amino acids m/tr, N = Nitrogen m/tr, Co = Coenzyme m/tr, Nu = Nucleotide m/tr, Ca = Carbohydrate m/tr, Li = Lipid m/tr, Ps = Polysaccharide m/tr, Ce = Cell envelope m/tr, 2M = Secondary metabolism, Rx = Redox, Tr = Transferases, Ot = Other enzymes. Where m/tr stands for "metabolism and trasportation" and p/c means "production and conversion".