Supporting Information

Maslov et al. 10.1073/pnas.0903206106

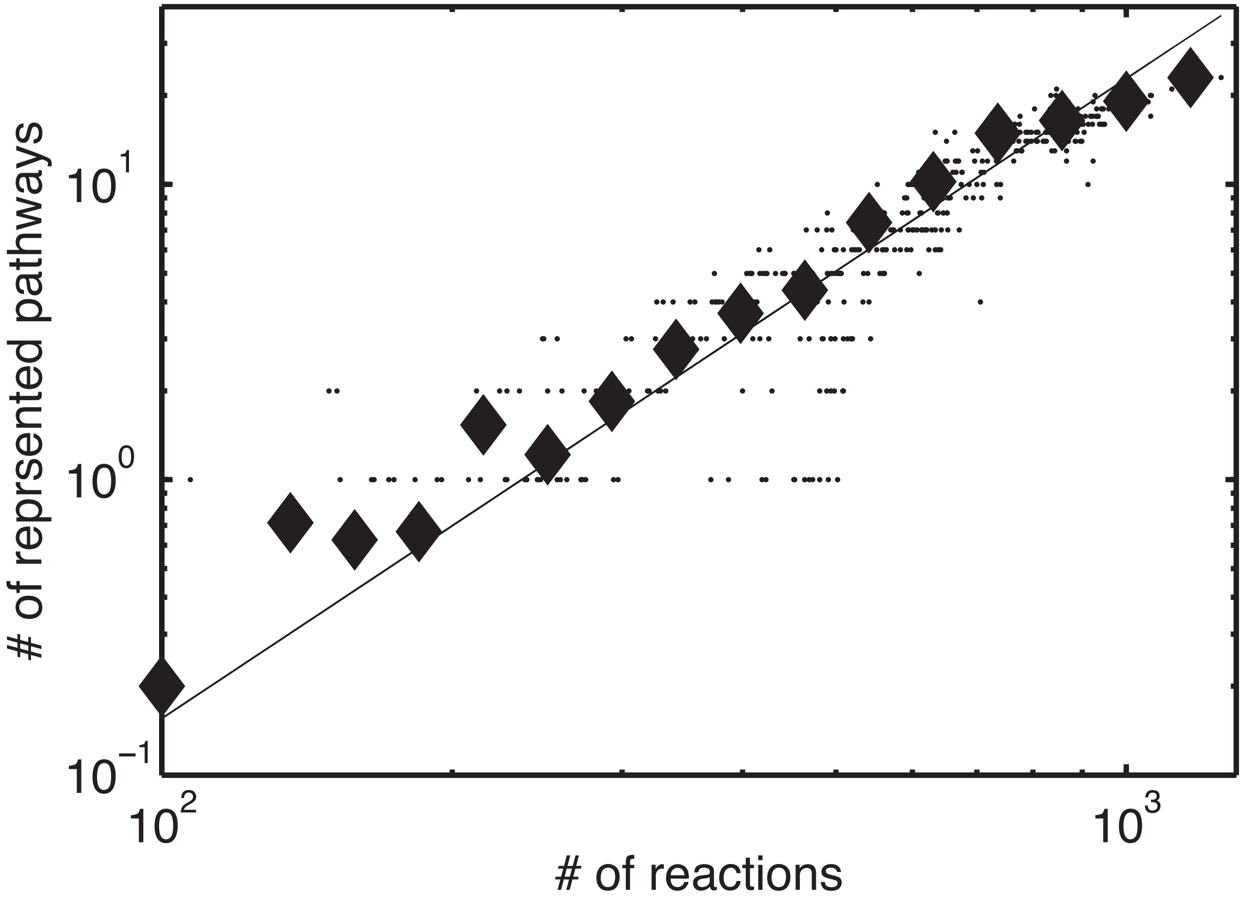

Fig. S1. The number of metabolic pathways in a prokaryotic genome scales faster than linearly with the number of reactions in its metabolic network. The dots represent 451 fully sequenced prokaryotes in the KEGG database, whereas filled diamonds are the same data logarithmically binned along the *x* axis. The mapping of reactions to known metabolic pathways is taken from the KEGG database (ftp.genome.jp/pub/kegg/ligand/reaction/reaction_mapformula.lst). A pathway is considered to be adequately represented in a genome if more than half of its reactions are present. The best power-law fit (solid line) has a slope 2.2 ± 0.2.

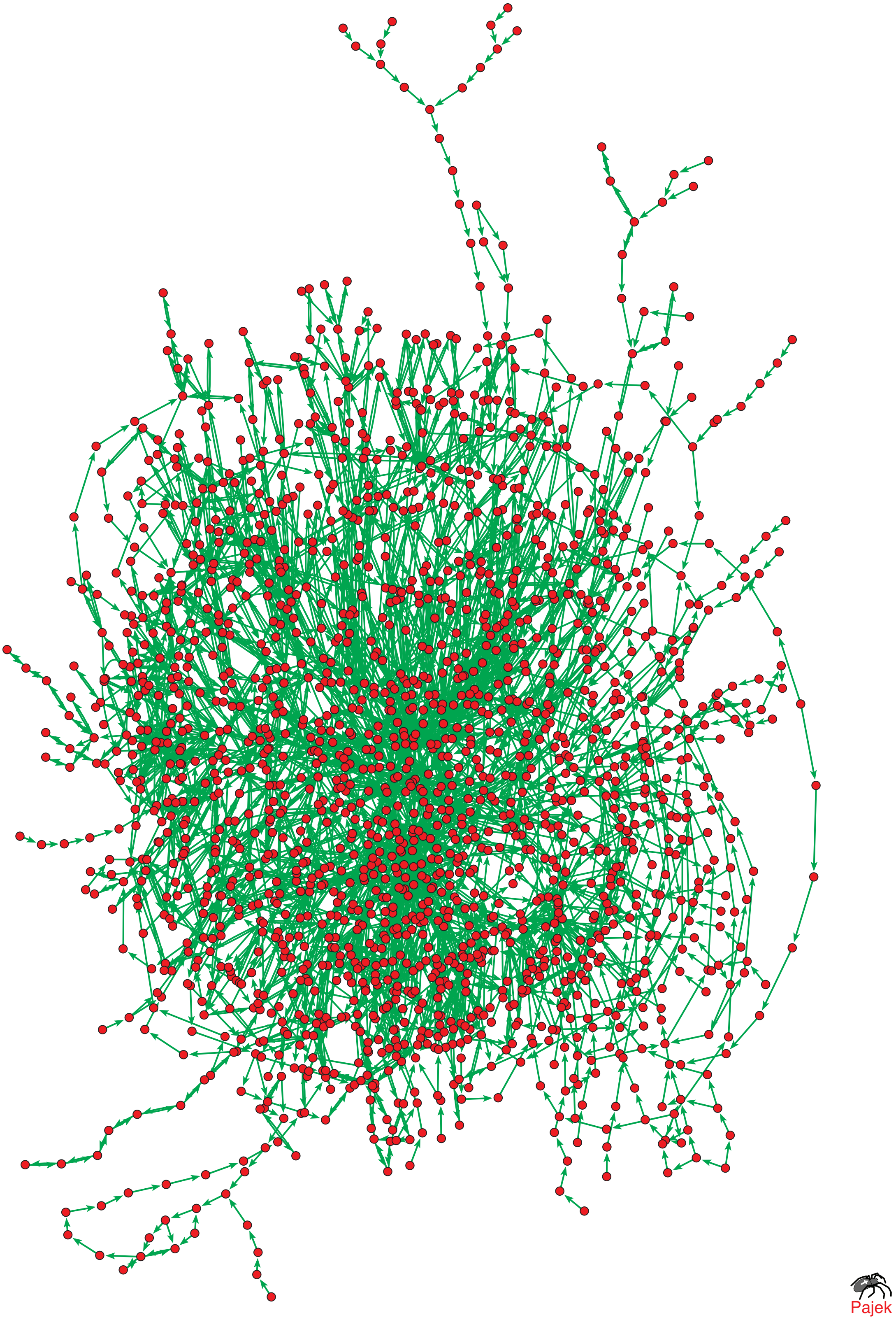

Fig. S2. The universal network used in our study formed by the union of all reactions listed in the KEGG database. The directionality of reactions and connected pairs of metabolites are inferred from the map version of the reaction formula: (ftp.genome.jp/pub/kegg/ligand/reaction/reaction_mapformula.lst). We then kept only the metabolites located upstream of the central metabolism. This left us with 1,813 metabolites connected by 3,714 edges.

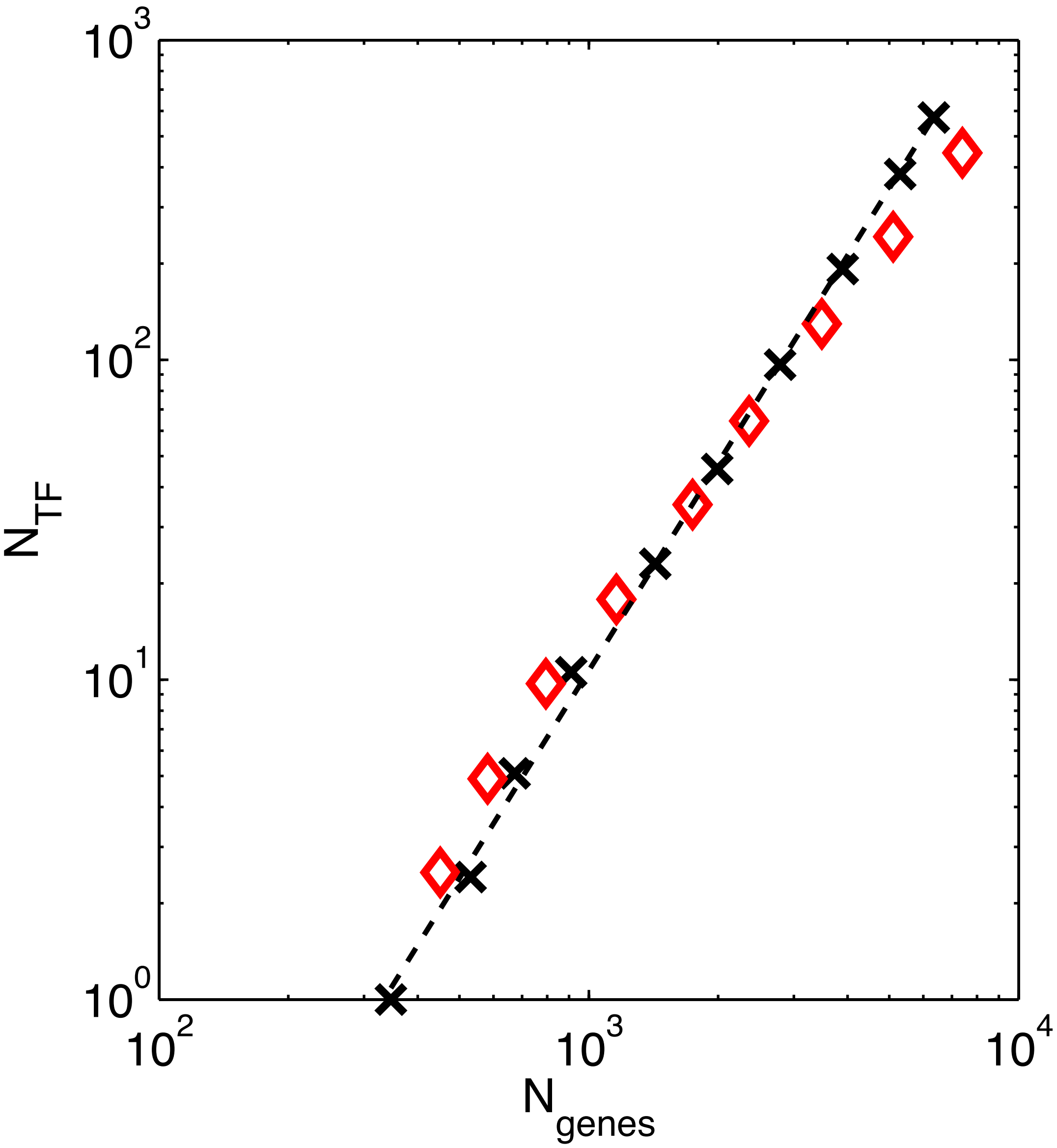

Fig. S3. Toolbox model when simulated on different universal networks of the same size ($N_{\mathrm{univ}} \approx$1,800) generates nearly identical N_{TF} vs. N_{met} plots. Red diamonds indicate the universal network made by the union of all KEGG reactions (same as red diamonds in Fig. 4*A*). Black Xs indicate the universal network formed by random walks on the fully connected graph. The data were log-binned for clarity.

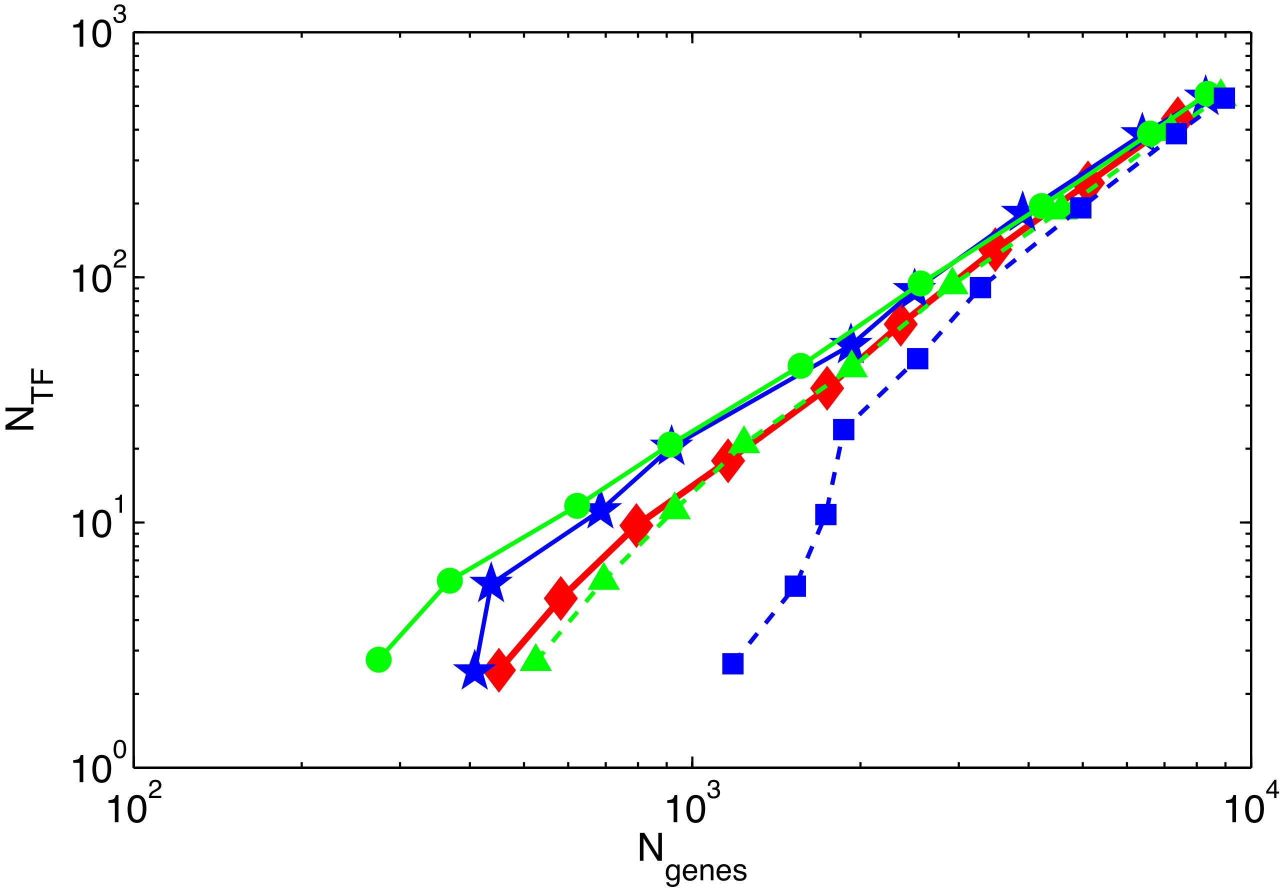

Fig. S4. Comparison of scaling in the standard toolbox model (red diamonds in this figure are the same as in Fig. 4*A*) with its variants in which lengths of attempted HGT pathways are drawn from a predetermined probability distribution $\pi(L)$. The rules of our model are modified as follows: a self-avoiding random path of length L drawn from $\pi(L)$ starts at a new nutrient/leave and follows edges of the universal network. If this new branch intersects the existing metabolic network, it is deemed evolutionary favorable, and its nodes starting from the leaf and ending at this intersection point are added to the network. In the opposite case, branches that failed to connect to the existing metabolic network and thus do not contribute to biomass production are discarded. Different symbols correspond to different functional forms of $\pi(L)$: an exponential $\pi(L) \approx \exp(-L/L_0)$ with $L_0 = 15$ (green circles and triangles) and a power law $\pi(L) \approx \mathrm{L}\hat{}^{-\delta}$, with $\delta = 1.5$ (blue stars and squares). In 2 of these models (green triangles and blue squares) we also introduced a delay in removal of redundant genes generated when a horizontally transferred pathway is longer than necessary to connect to the existing metabolic network. At each time step corresponding to addition/removal of a pathway we remove a fraction $r = 0.1$ of redundant genes. A typical redundant gene is thus likely to survive on average 10 pathway additions/removal steps.

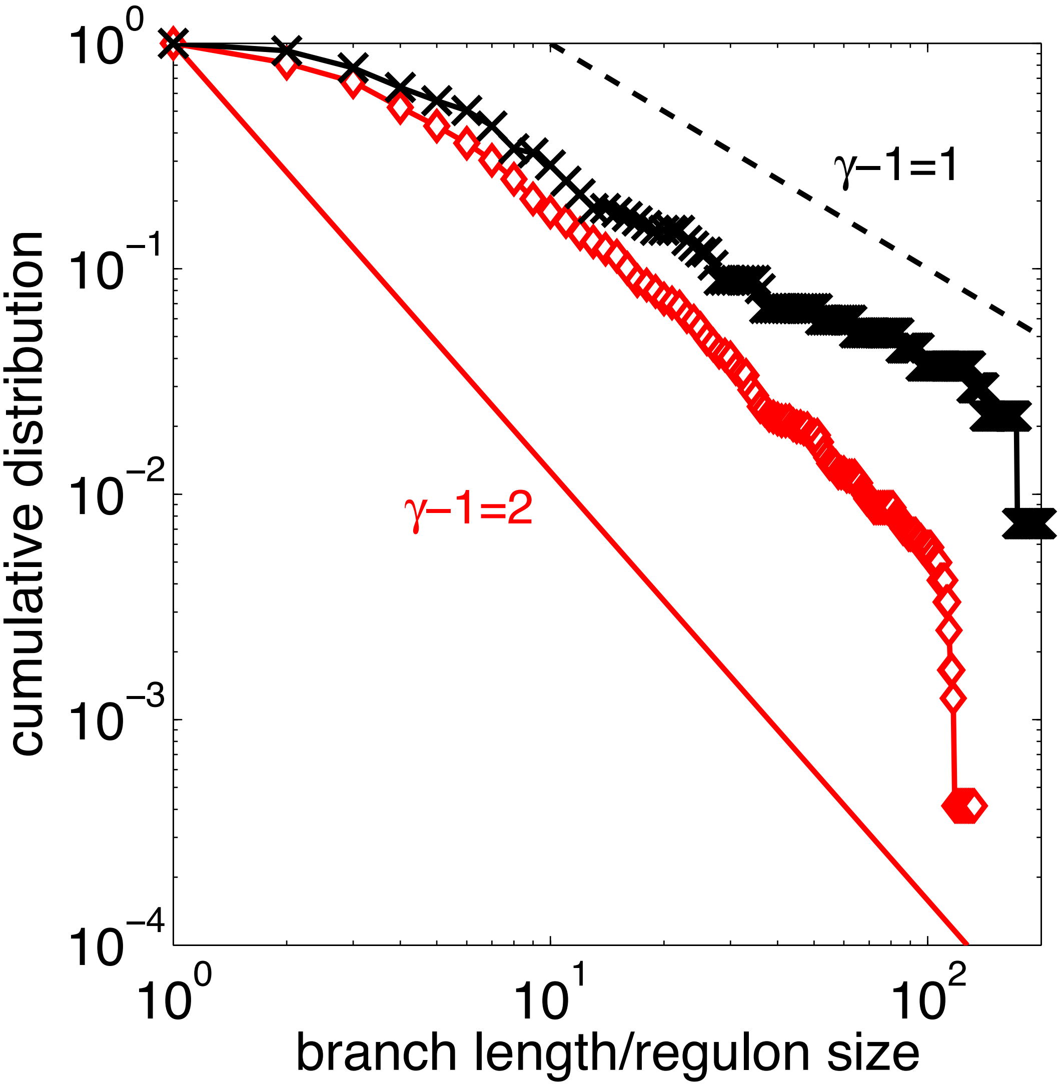

Fig. S5. The cumulative distribution of branch lengths in our model with $N_{\mathrm{met}} = 400$ simulated on the KEGG universal network ($N_{\mathrm{univ}} = 1{,}800$) (red diamonds) compared with the regulon size distribution in *E. coli* according to the regulon database (black Xs). One can see that the real-life regulon size distribution has longer tail than that of coregulated metabolic branches of our model in Fig. 2*A*.

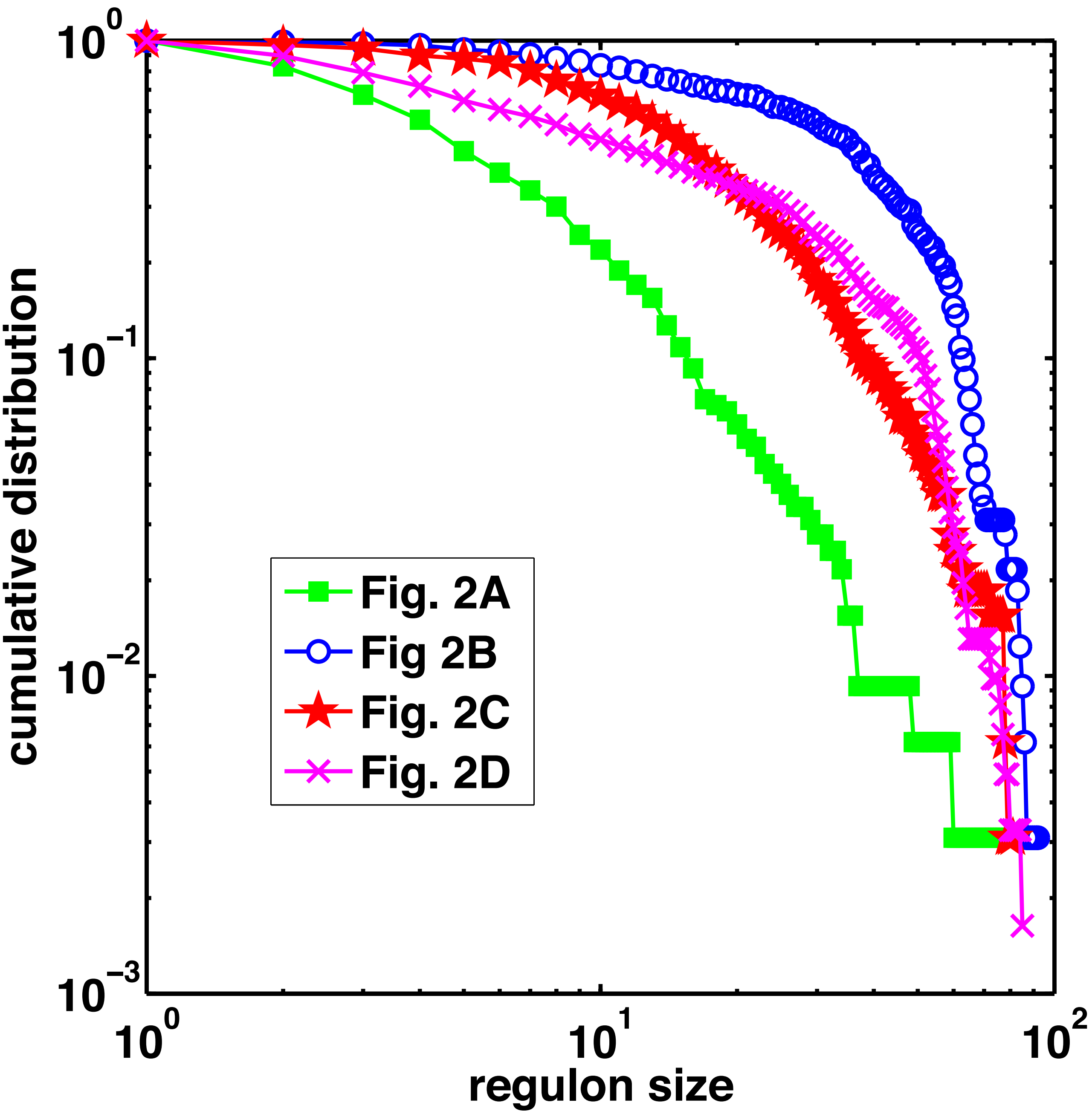

Fig. S6. Cumulative distributions of regulon sizes in our model shown in Fig. 2*A* (green squares), 2*B* (blue circles), 2*C* (red stars), and 2*D* (magenta Xs). Regulons in all models except for A are dominated by large hubs.

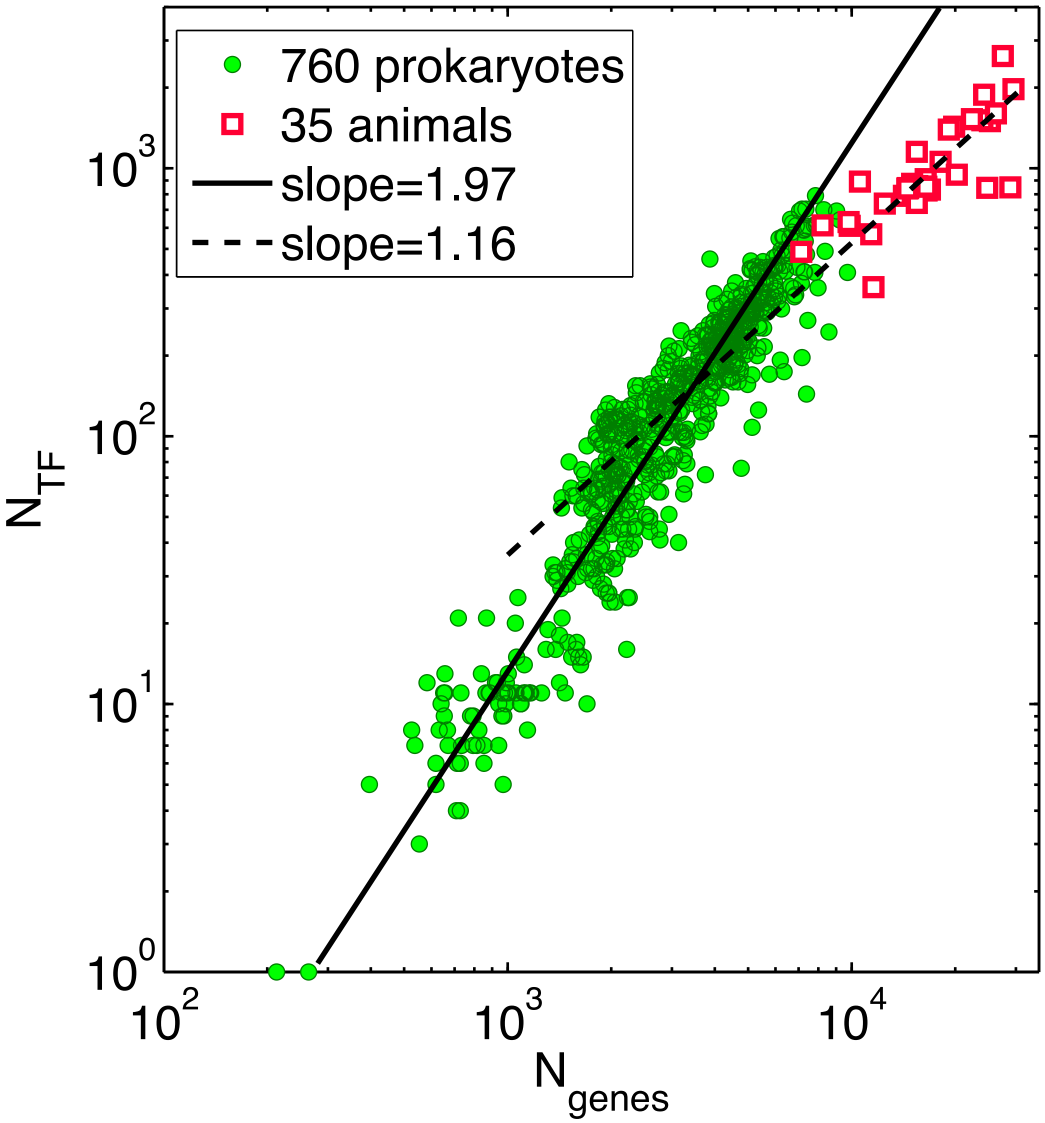

Fig. S7. The number of transcription factors plotted versus the total number of genes in 35 fully sequenced animal genomes in the KEGG database (red squares) compared with the same plot in 760 prokaryotic genomes (green circles as in Fig. 4*A*). The best-fit power-law exponents are 1.97 (solid line) and 1.16 (dashed line), correspondingly.

Other Supporting Information Files

Dataset S1 (XLS)