Supporting Information

Kreimer *et al.* 10.1073/pnas.0712149105

SI Text

Computing Modularity for Connected Network Analogs. Metabolic networks are often disconnected, comprising more than one connected component. As modularity is well defined for such disconnected networks, the analysis presented in the main text involved the straightforward application of the Newman algorithm to each species' network. To verify that our results are not biased because of the network's nonconnected structure, we repeated the modularity computation (and the ensuing analysis) using the following procedure: We divided each network into its connected components and removed all singleton nodes. We then aimed to create the closest connected network analog to the remaining network by adding the minimal number of random connections between its components for it to become connected. The overall correlation observed between the modularity scores of the original networks and their arbitrarily connected analogs was fairly high: $r = 0.5661$ ($P = 0$, Spearman rank correlation). We then proceeded to repeat the analysis presented in the article using the modularity scores of thearbitrarily connected networks. The results obtained were in accordance with the results reported originally, as follows: The correlation between the size of the network and its modularity is $r = 0.557$ ($P < 10^{-300}$, Spearman rank-correlation test), the correlation between modularity and mean nodes degree is $r = 0.3554$ ($P = 4.15 \cdot 10^{-11}$, Spearman rank-correlation test), and the correlation between modularity and network centrality is $r = 0.4697$ ($P < 10^{-300}$, Spearman rank correlation). The Spearman rank-correlation test between the extent of HGT and modularity in the 94 organisms yields a correlation of $r = 0.439$ ($P = 9.57 \cdot 10^{-6}$, Spearman rank correlation). Furthermore, host-associated bacteria have significantly lower modularity values than organisms in multiple ($P = 7 \cdot 10^{-4}$), aquatic ($P = 0.0013$), and terrestrial ($P = 0.0014$) environments, and facultative bacteria have lower modularity scores than aerobic bacteria ($P = 5 \cdot 10^{-3}$) (after correcting for multiple hypotheses testing using the Bonferroni correction). As evident, most of these results point to even stronger-magnitude correlations than those reported with the original modularity scores in the main text.

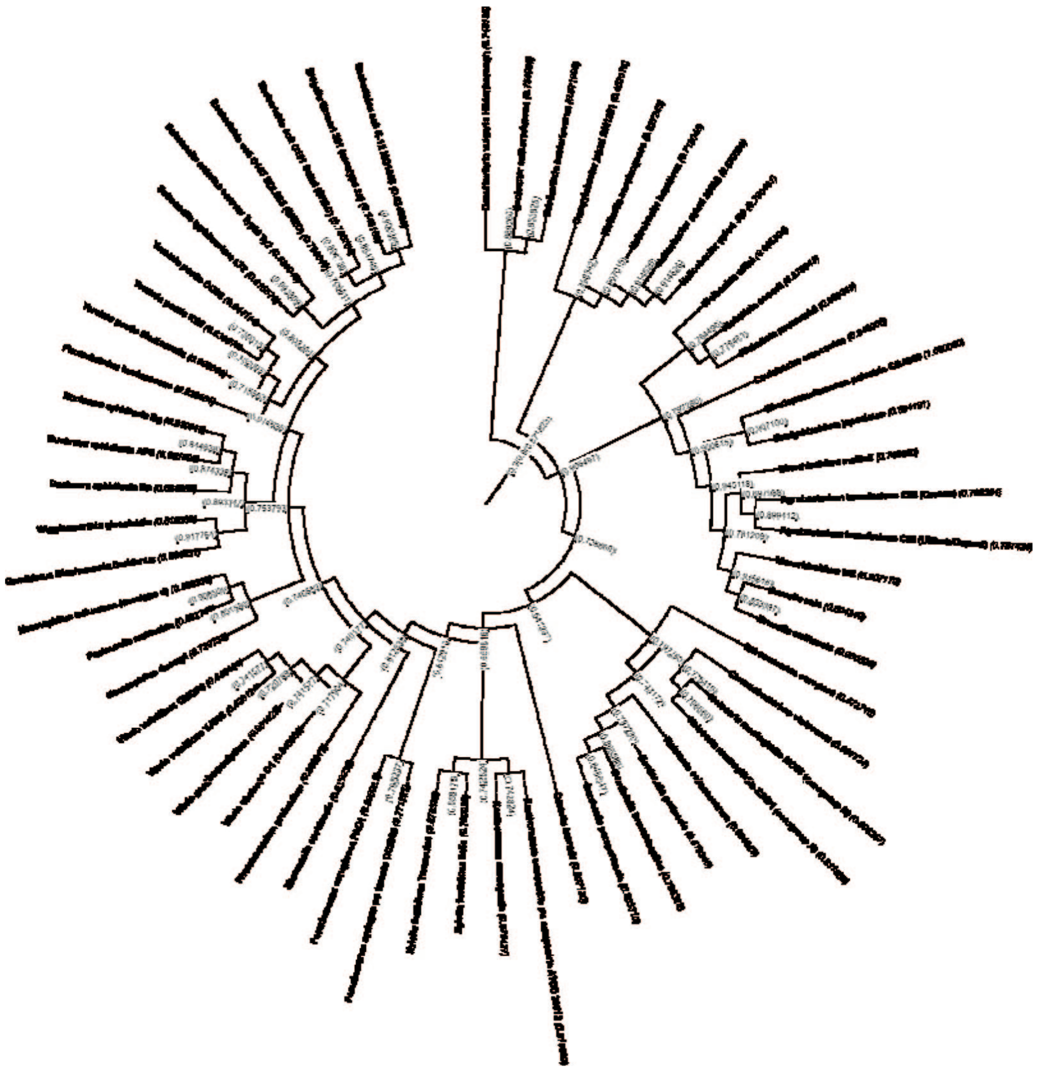

Fig. S1. Metabolic modularity scores across the phylum Proteobacteria. Numbers in parentheses represent modularity scores of the extant species (leaves) and their reconstructed ancestors (internal nodes).

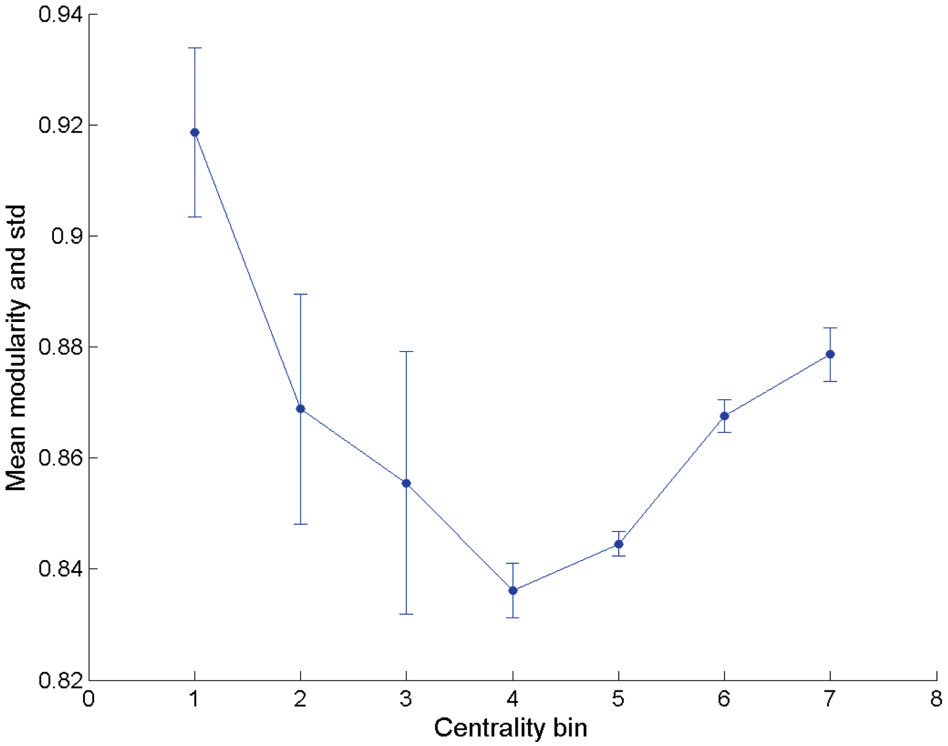

Fig. S2. The effect of loss of nodes of varying centrality on modularity. The simulation was carried out on the well defined metabolic network of *E. coli* (370 nodes, modularity score 0.8826). The nodes in this network were divided into seven equally sized bins determined by the centrality of each node, where 1 denotes the most central nodes, and 7 denotes the most peripheral nodes. The centrality of a node was defined as the average distance of a node from all other nodes in the network (*Methods*). For each bin, we ran 1,000 simulations where 25 randomly picked nodes from that bin were removed, and the modularity of the network after these nodes were removed was calculated. The figure depicts the mean and standard deviation of the obtained modularity scores for each bin.

Other Supporting Information Files

Dataset S1