Supporting Information

Ruths and Nakhleh 10.1073/pnas.1217630110

SI Text

1. Calculating Binding-Site Gain and Loss Rates from IUPAC Sequences

For each transcription factor (TF) sequence motif, provided in International Union of Pure and Applied Chemistry (IUPAC) code, we calculated their spontaneous gain and loss rates. Standard IUPAC nucleotide code describes ambiguous sites in a sequence motif, where a single IUPAC character may represent more than 1 nt. The gain rate is calculated as follows: Given a base pair mutation in the sequence, what is the probability that the sequence deviates from the IUPAC sequence at one site and that the mutation mutates it to a sequence that matches its motif? Let M be the IUPAC code for a TF motif of length N and let $M[i]$, for $1 \le i \le N$, denote the character at position i in the sequence M. We assume a simple model where all nucleotides and mutations between nucleotides are equally likely. We define a score function that returns the number of ambiguous nucleotides for each IUPAC character:

$$\mathrm{c}(x) = \begin{cases} 1, & x \in \{A, C, T, G\} \\ 2, & x \in \{M, R, W, S, Y, K\} \\ 3, & x \in \{B, D, H, V\} \\ 4, & x \in \{N, .\}. \end{cases}$$

Provided the score of an IUPAC character x, the probability that one site matches it is $c(x)/4$ and the probability that the site does not match is $1 - c(x)/4$. The probability that the sequence deviates from the sequence at only the first site (in the equation, $M[1]$) is

$$\frac{4 - c(M[1])}{4} \prod_{j \neq 1}^{N} c(M[j])/4.$$

Then, provided we know a mutation occurs at the first site, the probability that it mutates to a nucleotide that matches the IUPAC character is $c(x)/3$. So, the probability that the sequence deviates at only one site and that that site mutates to a matching nucleotide is

$$\mathrm{gain}(M) = \frac{1}{N} \sum_{i}^{N} \frac{c(M[i])}{3} \frac{4 - c(M[i])}{4} \prod_{j \neq i}^{N} c(M[j])/4.$$

The loss rate is calculated as follows: Given a base pair mutation within a binding site, the probability that the base pair mutation dissolves agreement with the TF motif is

$$\mathrm{loss}(M) = \frac{1}{N} \sum_{i}^{N} \frac{4 - c(M[i])}{3}.$$

Because we know the mutation occurs in the binding site, all base pairs match their corresponding IUPAC character. The probability that the mutation occurs at site i is $1/N$, and the probability that the mutation causes the nucleotide to not match the IUPAC sequence character x is $(4 - c(x))/3$.

Because both loss$(\cdot)$ and gain$(\cdot)$ are probabilities conditional on a base pair mutation, multiplying by the base pair mutation rate u gives the individual loss and gain rates for each TF motif M: $u\mathrm{loss}(M)$ and $u\mathrm{gain}(M)$.

2. Alternate Simulation Scenarios

This study leveraged carefully parameterized simulations to understand neutral patterns that may arise in the *Escherichia coli cis*-regulatory network over evolutionary timescales. Although all parameters in the main study derive from empirical values of the *E. coli* genome and regulome, a critical question remains as to the effect—whether weak or strong—that deviations in these parameters may have on the results. Because the empirical parameters, like binding-site size, may present small or biased samples, it is important to understand the strength of the results provided noise in the parameterization. Specifically, we measure the effect of promoter heterogeneity, initialization condition, binding-site length, and population size on the results of the study.

In the following sections, we describe the other simulations we ran and discuss their results. Figs. S5–S10 provide an all-against-all comparison of the different experiments, split out by system, subgraph, and operon level.

Across all these experiments, we found that the results from the main study not only are robust to perturbations in the parameter values or starting conditions but also modify the results in a systematic fashion. For instance, decreasing the binding-site length increases the number of edges at equilibrium, which elevates the clustering coefficient, and does not significantly affect the results in the subgraph or operon level. In fact, we find that "small" changes in the parameters still yield results wherein feed-forward loops (FFLs) and other previously identified "adaptive" properties are not statistically significant from the null model. Only when we use homogeneous values for promoter lengths and mutation rates or use random walks instead of population genetic simulations do we notice drastically different results. Thus, the use of actual distributions to parameterize the null model—despite the minor differences in which these empirical distributions are derived from sequence-level data—is a significant and unique contribution to the identification of neutral patterns in the *E. coli* regulatory network.

Empirical Parameterization. As described in the main text, we parameterized the evolutionary simulations with empirical distributions of

- promoter length: inferred from coding regions on the *E. coli* genome;
- binding-site IUPAC sequences: downloaded from RegulonDB (1);
- average binding-site length: the average length of binding-site IUPAC sequences; and
- population size: estimated in previous studies.

These empirical distributions were retrieved from RegulonDB, which is kept up-to-date with the latest findings on the *E. coli* regulatory network. The population size for *E. coli* is a rough estimate for bacterial populations. For each operon, the results of the main study including the parameters used—promoter length and PWM—are provided in the table of operon results (Dataset S1).

Homogeneous Promoter Lengths and Mutation Rates. In this experiment, we investigated the significance of using actual distributions of promoter lengths and binding-site gain and loss rates derived from PWMs by comparing them to evolutionary simulations parameterized with average values only. To test the effect of using a homogeneous promoter length, used in previous studies like ref. 2, we parameterized each operon with the same promoter length, set to the average of the actual distribution in *E. coli*. In a similar manner, we replaced the heterogeneous distribution of binding-site gain and loss rates—derived from their respective IUPAC sequences—with the average gain and loss rate. All other parameters

were left unchanged, including the average binding-site length (20 bp) and population size (10^9).

Under this model, all of the results of the main study are effectively annulled. On the system level, the homogenous null model expects 5,232 interactions (z-score = −71.4) and an average clustering coefficient of 0.552 (z-score = −14.6), and the in-degree and out-degree distributions were significantly different. The distributions of promoter length and in-degree differed by 97%. Among many other differences in the subgraph distribution, the bifan and feed-forward loop subgraphs had z-scores 10 times and 3 times larger than the nonhomogeneous model. A bifan is a directed graph on four nodes, two of which are designated target genes and each of the other two is designated as a regulator of both target genes. At the operon level, the majority of operons fell significantly outside neutral expectations.

Alternate Initial Condition. In the main study we seeded the simulations with a random initial network, such that all TF operons are autoregulatory and any non-TF–encoding operon is regulated by a random TF. Such a configuration guarantees a minimally viable network such that the loss of any binding site would render the network nonviable. Further, this initial condition has no "interesting" motifs besides a single input module, because no TFs can regulate another TF, so we do not bias the subgraph results with the initial random network. However, there are other methods for generating such random networks. To test the effect of the initial conditions on the results of the main study, we defined an alternative random initial network strategy such that both non-TF– and TF-encoding operons are regulated by a randomly chosen TF. Under this strategy, we do not force TFs to be autoregulatory, but all operons still have only one incoming regulatory interaction. So, although there may be linear pathways and feedback loops in this initial random network, few other subgraphs, like FFL or bifan, exist in the network. In fact, this bias is present in the results: The second most common subgraph, a three-node linear pathway, is much more common in the alternate initial condition than in the main study; however, z-scores for all other motifs are nearly identical.

Under this experiment, the results of the main study were not only repeated and validated, but also the values themselves were closely matched. Therefore, the topology of the initial condition, evidenced by the agreement of the results from these two different approaches, does not have a significant effect on the results of the main study.

Shorter Binding-Site Length. Although the average binding-site length is calculated from actual *E. coli* IUPAC sequences, the length of these sequences spans a broad distribution. Consequently, the average used in the main study (20 bp; Fig. S1) may be a poor approximation of the ground truth. Although the specific gain and loss rates of each binding motif are calculated from their IUPAC sequence, we simulate all binding sites as the same length, for computational purposes. Thus, by performing the same simulations with a shorter binding-site length (7 bp), we may examine the extent to which (*i*) the average binding site and (*ii*) the use of a homogeneous binding-site length may influence the simulations. In regard to the latter, the effect of using heterogeneous binding-site lengths could be extrapolated from the effect that changing the length from 20 bp to 7 bp has on the overall resulting networks.

Decreasing the binding-site length reduces the loss rate of binding sites overall, because smaller binding sites present a smaller target for point mutations. Consequently, using a 7-bp binding site increases the average number of edges in equilibrium to 1,127 (z-score = −3.77) from 1,039 (results from *Empirical Parameterization*) and the average clustering coefficient to 0.231 (z-score = −0.7) from 0.162. Despite these slightly elevated numbers, decreasing the binding site corroborated the results of the main study overall, especially including the subgraph- and operon-level results. Thus, the binding-site length plays a "minor" role in the results of the main study, and so we may expect that enhancing the simulator to support heterogeneous binding-site length would also result in minor changes to the overall results.

Smaller Population Size. Although the population size used in the main study is widely accepted as a reasonable effective population size for bacteria over the evolutionary timescale under investigation in this study, the role that the effective population size has in the null model is of significant interest to this study. To investigate the role of population size on the null model, we performed simulations with an effective population size of 10^6. Altering the population size affects the mechanics of genetic drift: Reducing the population size in fact makes genetic drift a more powerful evolutionary force (all else staying equal).

The smaller population size resulted in substantially (around 300) more edges in equilibrium, which in turn increased the clustering coefficient and degree distributions. Smaller populations allowed for more binding-site gains to occur in long promoter regions, increasing the number of regulatory edges in the network. However, even under the smaller population size, FFLs occurred with a z-score of 1.3, too low to reject the null model.

Random Walk. Although an evolutionary perspective is at the core of our study, a question arises as to whether full-blown population genetic simulations are required to reproduce the results. In fact, previous work that leveraged a neutral evolutionary approach to replicate clustering patterns in eukaryotic enhancers opted for random walks vs. population genetic simulations (3). Because random walks are much simpler from a computational and modeling perspective, they are indeed the preferred choice if they accurately represent the evolutionary dynamics of the study.

We simulated network evolution, using random walks instead of a population genetic lifecycle. Each random walk is initialized identically to the main study and each step in the random walk corresponds to a binding-site mutation (gain or loss). If the mutation results in a nonviable network, the step is rejected and a new mutation is sampled. This is the same process used in ref. 3, where they found "no appreciable difference" between the results of population genetic simulations and those of random walks. We measured the number of steps until an equilibrium in number of edges is reached (similar to the population genetic approach) and let that determine the number of steps in each walk.

Under random walks, the number of edges in equilibrium nearly doubles to 1,918 (z-score = −49) and the clustering coefficient also increases to 0.73 (z-score = −9). Like the smaller population size, there is a strong linear correlation between the length of a promoter and its in-degree; in both the main study and the real *E. coli* network this relationship is not so distinct. The degree distributions of the random walk networks are skewed to higher degrees. At the subgraph level, the subgraph distributions are indeed different from those produced by population genetic simulations (even under smaller population sizes), but FFLs, bifans, and single-input modules (SIMs) still occur at levels similar to the actual *E. coli* network, although still not at equivalent values to the other simulations.

These results show that random walks, where nonviable mutations can be resampled, do not accurately represent the random walk performed by a population guided under genetic drift. Intuitively, the more accessible areas will continue to be explored by a large population whereas inaccessible areas (high nonviability or against the "mutation bias") will not be explored. In this case, because there is a loss bias for binding sites, continuing to gain sites (and edges in the network) operates against the mutation bias present in the *E. coli* genome. Hence, population genetic simulations give us a different answer than random walks.

1. Gama-Castro S, et al. (2011) RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* 39(Database issue):D98–D105.
2. Lynch M (2007) The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* 8(10):803–813.
3. Lusk RW, Eisen MB (2010) Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genet* 6(1): e1000829.

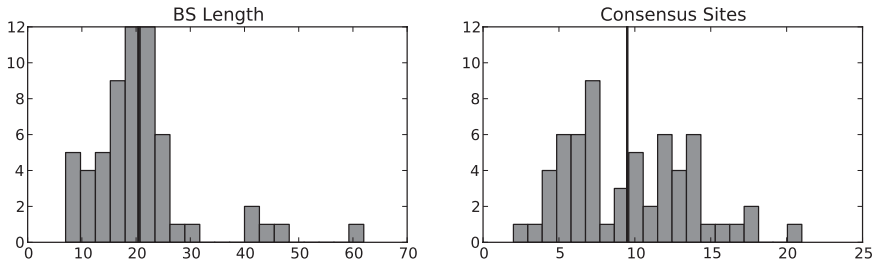

Fig. S1. Histograms of the binding-site length for the *E. coli* IUPAC sequences in this study are measured in total length (*Left*) and consensus length (*Right*). Total length is the length of the TF motif, including ambiguous sites. Consensus length is the number of nonambiguous sites. The average for each distribution is indicated with a vertical bar. The average binding-site length is 20 bp.

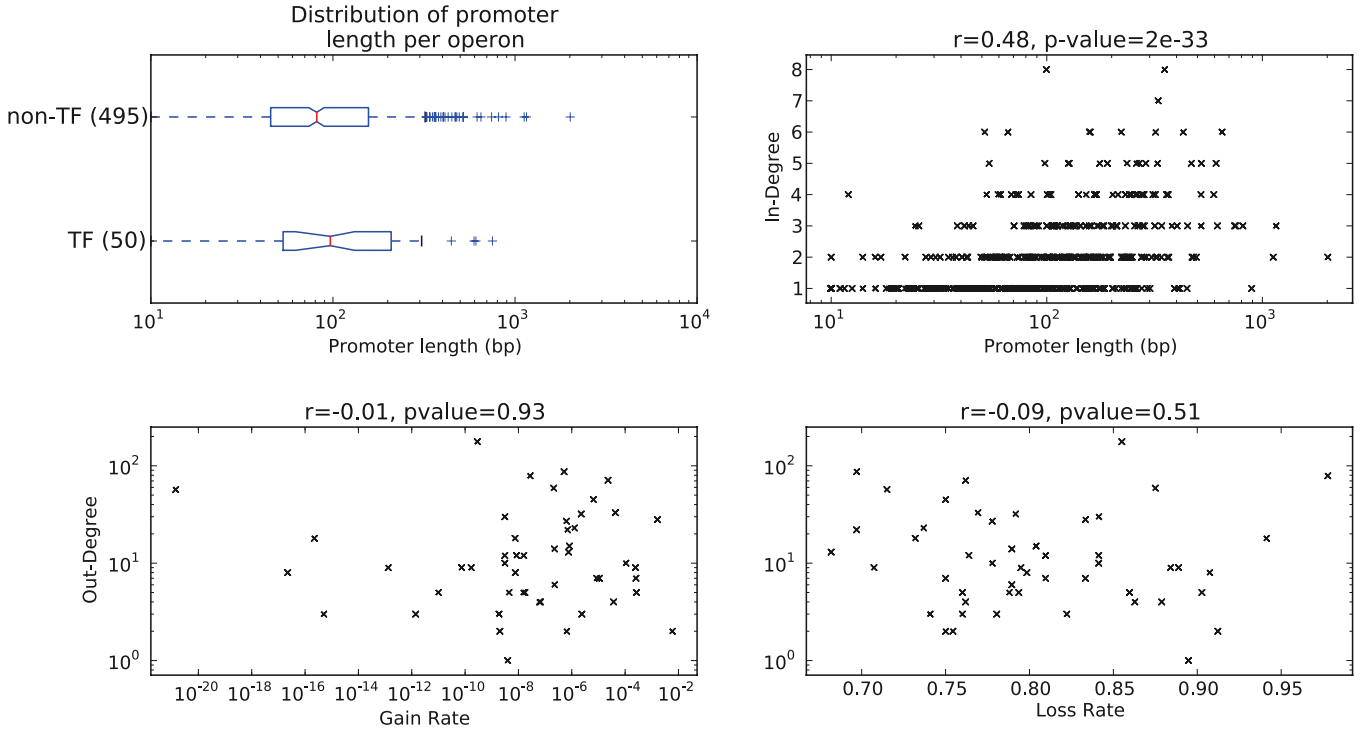

Fig. S2. Relevant distributions for genomic- and network-level properties of the *E. coli* regulatory network. (*Upper*) Information on distributions related to promoter length; (*Lower*) plots of binding-site gain and loss rates for each operon against their out-degree. *Upper Left* compares the distributions of promoter lengths of operons that encode a TF and that do not encode a TF ("non-TF"). The number of operons represented in each distribution is listed in parentheses, with all operons in the regulatory network accounted for. Although TF-encoding operons have an elevated average and median promoter length, it is not significant (Mann–Whitney nonparametric test, *P* value = 0.18). (*Upper Right*) Each operon is plotted with its in-degree and promoter length; we report the Pearson correlation coefficient and significance above the plot. (*Lower*) The out-degree of TF-encoding operons is plotted against their spontaneous gain rate (*Left*) and loss rate (*Right*), as computed from their position weight matrices.

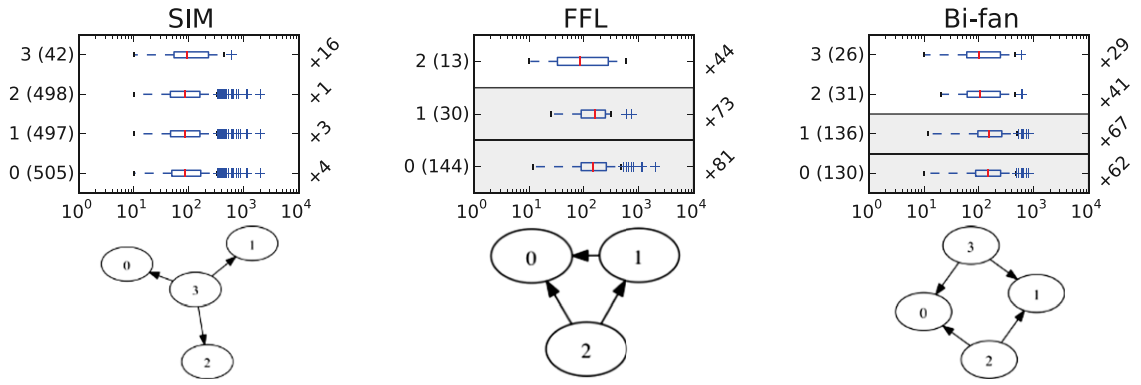

Fig. S3. Distributions of promoter length for operons that participate in single-input modules (SIM), feed-forward loops (FFL), and bifans are presented as boxplots per node. The left axis provides the node label, which corresponds to the node in the subgraph diagram (e.g., 0 or 1), along with the number of distinct operons represented in that distribution. The difference between the average promoter length in the node distribution and the average promoter length in the network is listed on the right axis. Distributions with significant uplift, assessed using a nonparametric Wilcoxon's rank-sums test, are indicated with a gray background behind the boxplot. The *P* values for these distributions are FFL0 = 4×10^{-11}, FFL1 = 5×10^{-4}, Bi − fan0 = 9×10^{-10}, and Bi − fan1 = 8×10^{-11}.

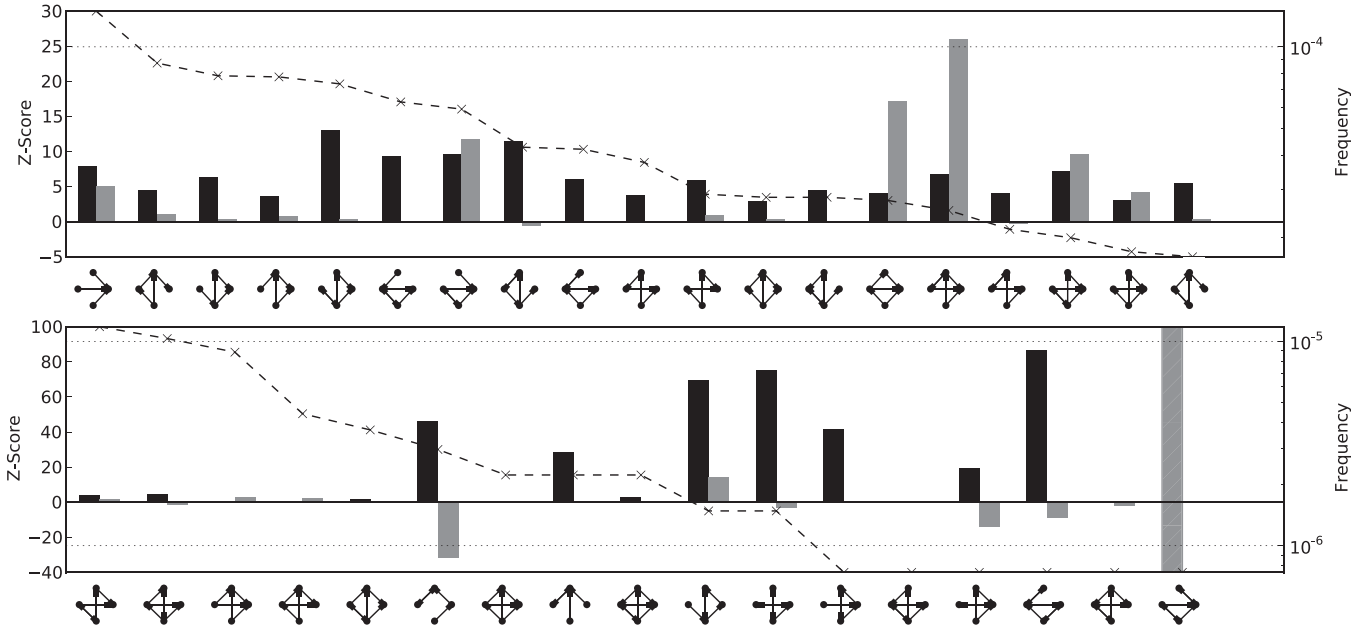

Fig. S4. The *z*-scores for low-frequency four-node subgraphs are ranked according to frequency in the *E. coli* network. The left axis provides the scale for the *z*-score (bars) and the right axis measures the frequency (dashed line) of each subgraph. For each subgraph, the *z*-scores for our model (black) and the edge-switching model (gray) are graphed side-by-side.

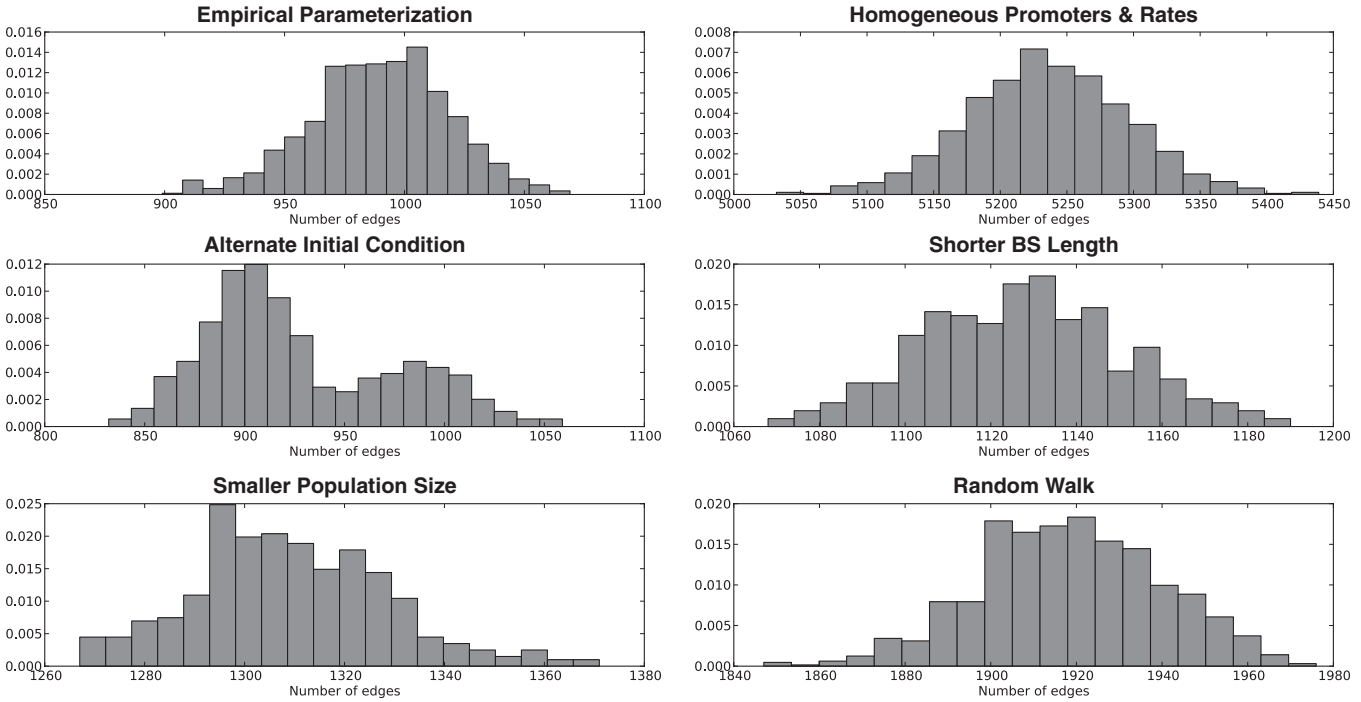

Fig. S5. Distribution of the number of edges at equilibrium under the alternate simulation scenarios.

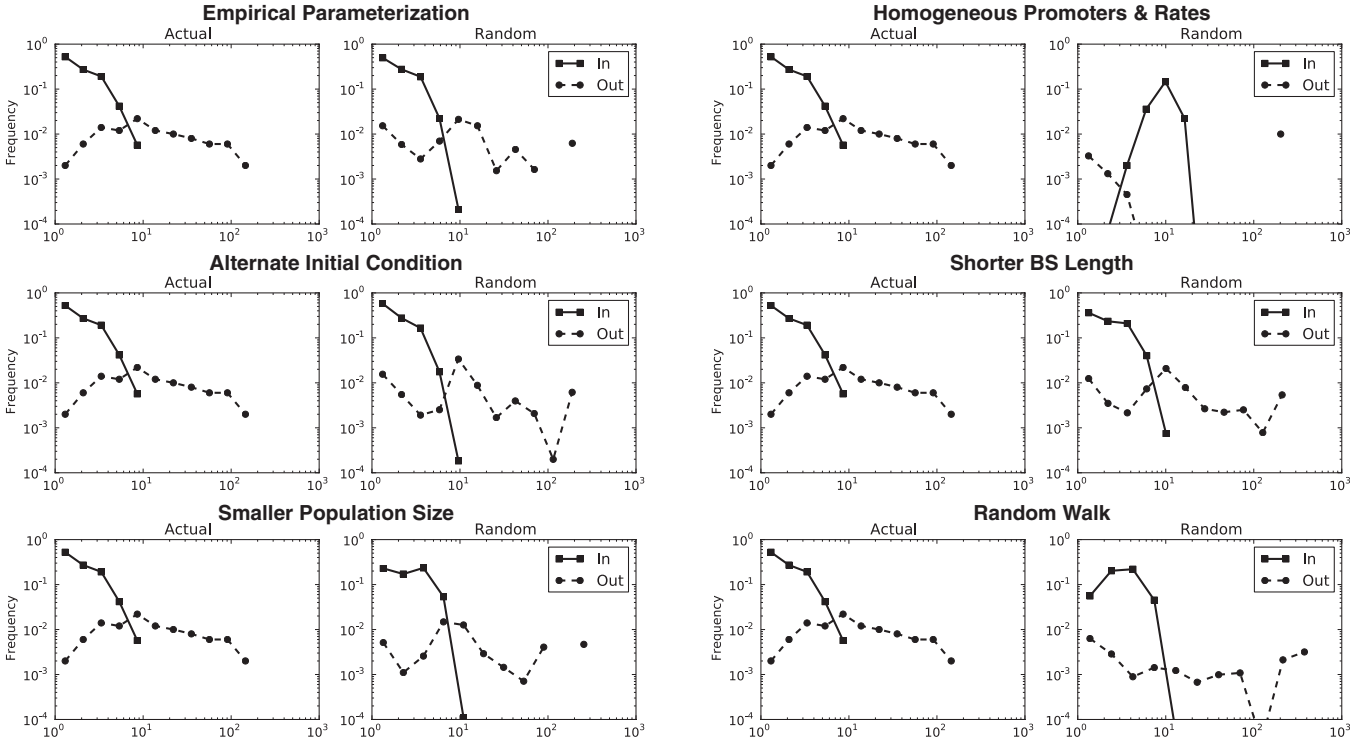

Fig. S6. Degree distributions—comparing actual to random—for the alternate simulation scenarios.

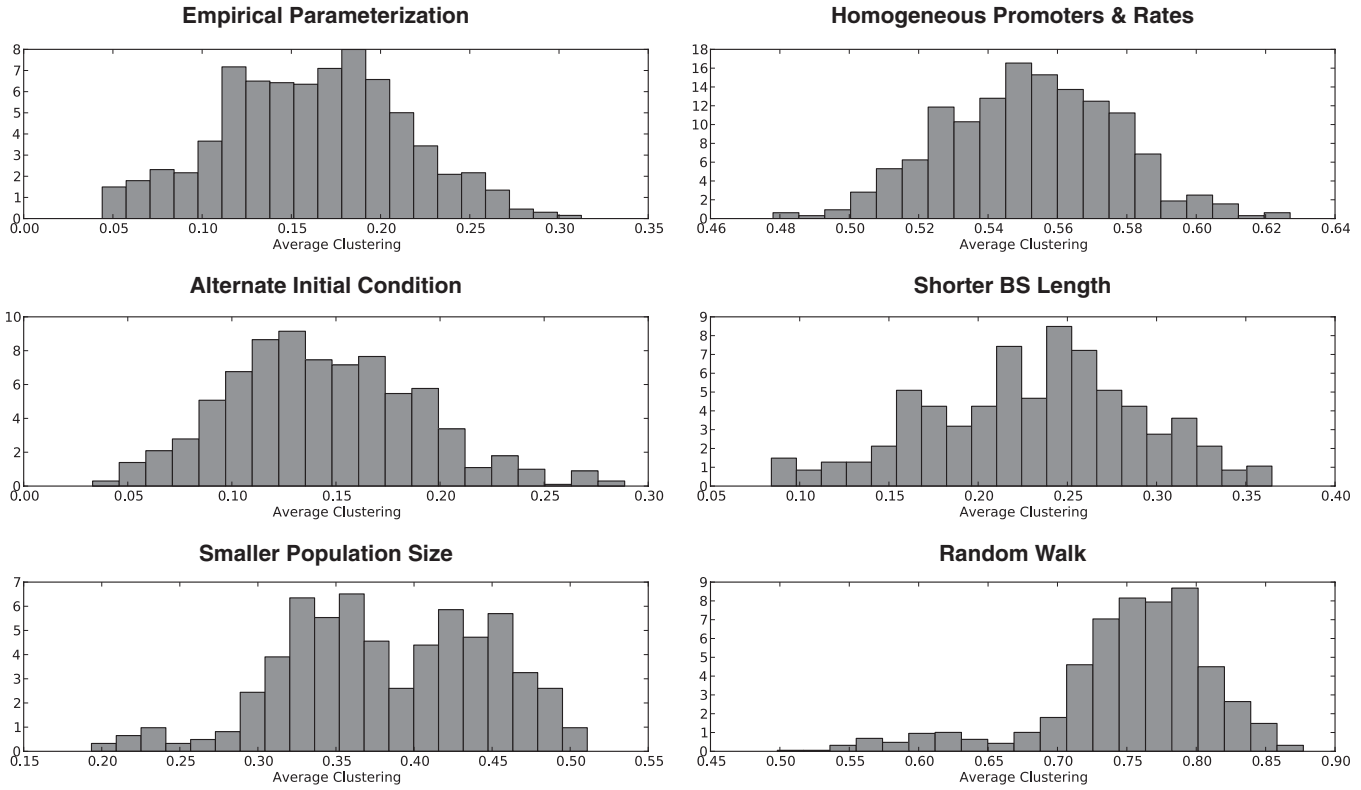

Fig. S7. Distribution of clustering coefficients under the alternate simulation scenarios.

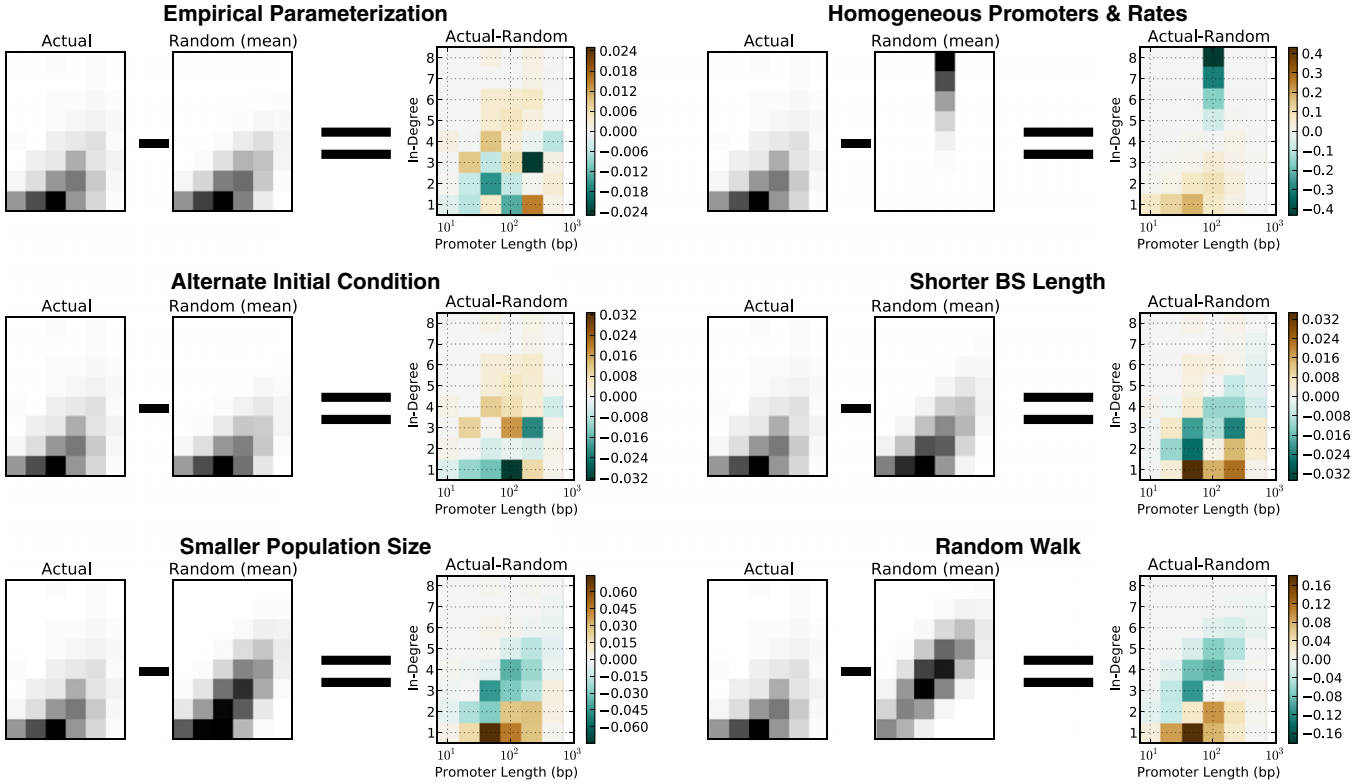

Fig. S8. Joint distribution of in-degree and promoter length—comparing actual to random—under the alternate simulation scenarios.

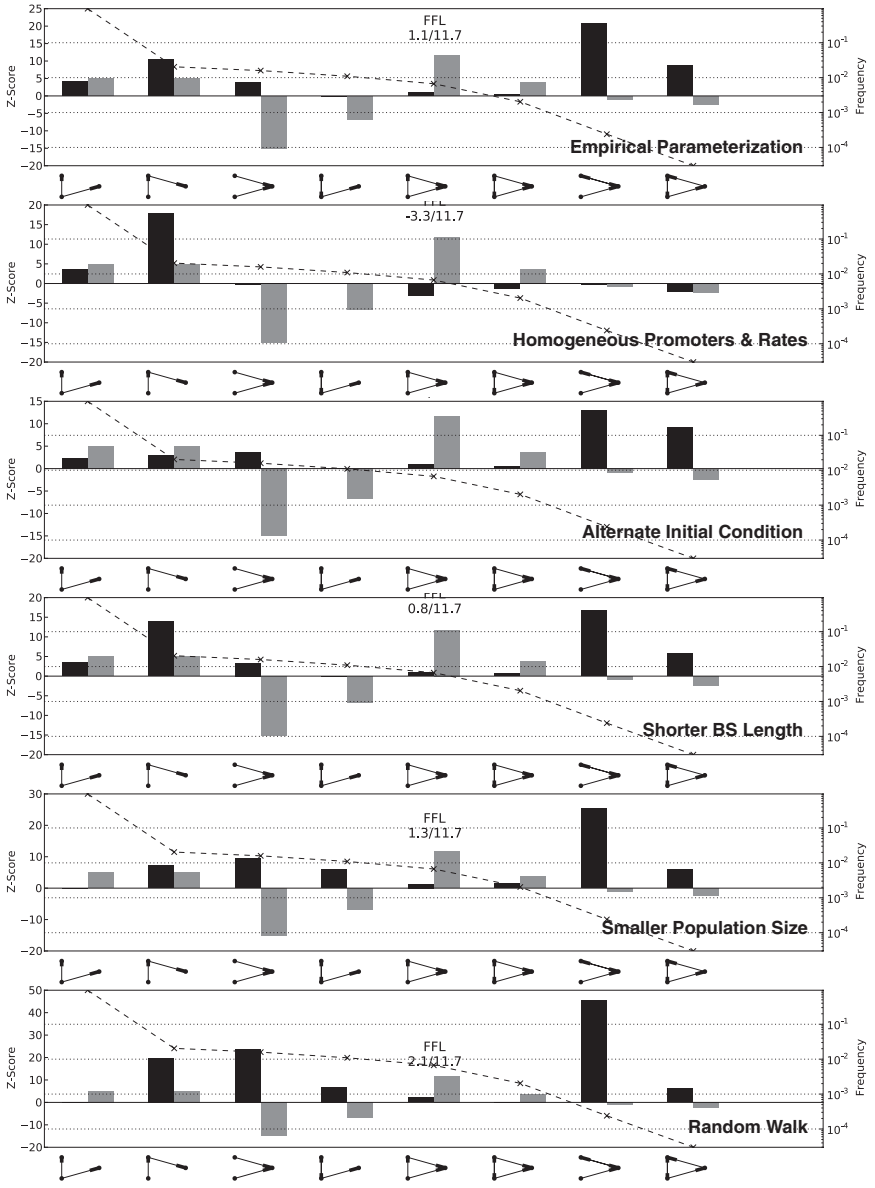

Fig. S9. The *z*-scores for three-node subgraphs under the alternate simulation scenarios, sorted by the frequency in the *E. coli* network.

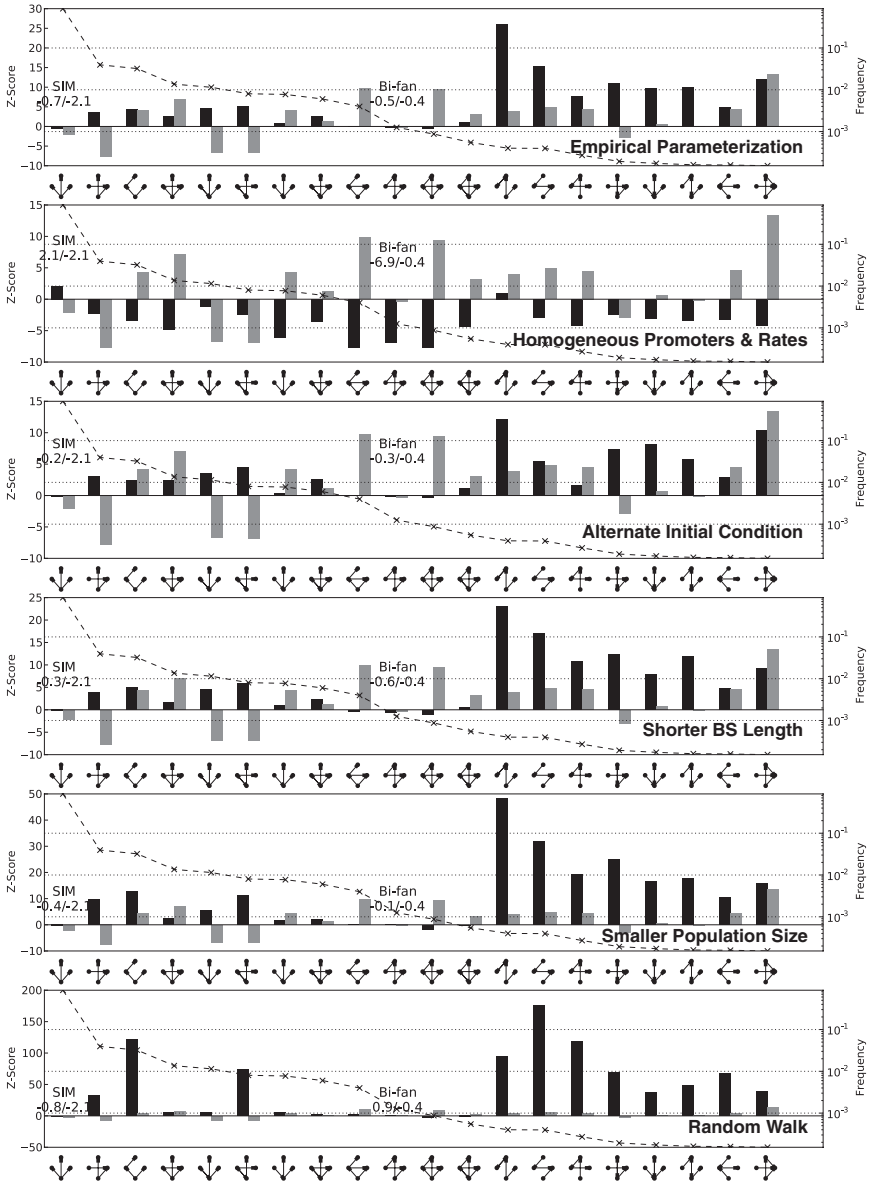

Fig. S10. The *z*-scores for the most common four-node subgraphs under the alternate simulation scenarios, sorted by the frequency in the *E. coli* network.

Other Supporting Information Files

Dataset S1 (XLS)