Computational design of genomic transcriptional networks with adaptation to varying environments

Javier Carrera[a,b], Santiago F. Elena[b,c], and Alfonso Jaramillo[a,1]

[a]Synth-Bio Group, Institute of Systems and Synthetic Biology, Universite d'Evry Val d'Essonne, Genopole®, Centre National de la Recherche Scientifique (CNRS UPS3509), 91030 Evry Cedex, France; [b]Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas-UPV, 46022 València, Spain; and [c]The Santa Fe Institute, Santa Fe, NM 87501

Edited by* Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved July 2, 2012 (received for review January 6, 2012)

Transcriptional profiling has been widely used as a tool for unveiling the coregulations of genes in response to genetic and environmental perturbations. These coregulations have been used, in a few instances, to infer global transcriptional regulatory models. Here, using the large amount of transcriptomic information available for the bacterium *Escherichia coli*, we seek to understand the design principles determining the regulation of its transcriptome. Combining transcriptomic and signaling data, we develop an evolutionary computational procedure that allows obtaining alternative genomic transcriptional regulatory network (GTRN) that still maintains its adaptability to dynamic environments. We apply our methodology to an *E. coli* GTRN and show that it could be rewired to simpler transcriptional regulatory structures. These rewired GTRNs still maintain the global physiological response to fluctuating environments. Rewired GTRNs contain 73% fewer regulated operons. Genes with similar functions and coordinated patterns of expression across environments are clustered into longer regulated operons. These synthetic GTRNs are more sensitive and show a more robust response to challenging environments. This result illustrates that the natural configuration of *E. coli* GTRN does not necessarily result from selection for robustness to environmental perturbations, but that evolutionary contingencies may have been important as well. We also discuss the limitations of our methodology in the context of the demand theory. Our procedure will be useful as a novel way to analyze global transcription regulation networks and in synthetic biology for the de novo design of genomes.

automated design | synthetic genomics | genome refactoring | evolutionary computation

Organisms have evolved mechanisms for regulating transcription to better adapt to changing environments. Could such regulation be engineered in a different way (1, 2)? Recent experiments investigating the evolvability of bacterial transcriptional regulatory networks (TRNs) have shown that the massive addition of new links to the network does not significantly alter cell growth. Isalan et al. (3) added transcriptional fusions of promoters with different master transcriptional regulators and showed that *Escherichia coli* (*E. coli*) tolerated almost all rewired networks; however, growth was perturbed by as much as 5% (3). This inherent predisposition of *E. coli* networks to dampen extreme changes in their circuitry enables the possibility of conducting genome-wide rewiring (4). Global transcription regulation could also be analyzed by comparing the regulatory models from distant organisms, provided they show a similar response to the set of studied environments. In this way, they could provide alternative regulatory models, although the lack of knowledge of species-specific selective pressures may blur the conclusions. We will propose here an alternative evolution experiment, which will be conducted computationally thanks to the availability of a quantitative model for the genomic transcriptional regulatory network (GTRN) of *E. coli*.

Global models of transcription regulation are essential to understand the function of an organism in alternative environments. The analysis of the structure of GTRN has unveiled many design principles, such as the identification of local patterns of regulation with defined function (5). How predictable should a model be in order to be able to evolve a global TRN? The relationship between network structure and function is best described by models based on ordinary differential equations (ODEs) that implement instances of the regulatory network. Monitoring of gene expression at a genome-wide scale allows assigning parameter values to global models of transcription regulation (6). If it were possible to create an ODE model for the global transcriptional regulation and signaling of a given genome, then we would be able to predict the function of a network even after rewiring it in silico, allowing the generation of alternative models with similar behavior. We will show that this can be done by adapting an existing ODE model for the TRN of *E. coli* (7) to include the required signal transduction. The evolutionary computational methodology here proposed is general, and it could be used with other ODE models for TRNs (8–13).

The computational design of small TRNs was first proposed by using computational evolution with a system of ODEs describing the TRN (14), although no nucleotide sequence was generated for the evolved TRN. Recently, the use of a modular approach based on the assembly of biological part models has allowed the assignation of nucleotide sequences to the evolved TRN (15), which opened the door to the automatic design of genomic-sized sequences. For genomic-scale TRNs, we could take advantage of the available high-throughput functional genomics data to infer the required ODE models (7). Evolutionary TRN optimization requires defining a fitness function. A simple fitness function could be defined based on the expression levels of some selected genes. Alternatively, a more complex fitness function could be defined by linking gene expression to cell growth, which would allow evolving whole genome TRNs. We call this a GTRN, defined as a TRN (including signaling) together with a fitness function accounting for cell growth. It has recently been shown that the transcriptomic expression profiles are good predictors for instantaneous cell growth in *Saccharomyces cerevisiae* (16). Assuming that this relationship is true for other organisms, it can be hypothesized that the expression profile of a given system determines cell growth. This can also be rationalized by arguing that natural selection results in nearly optimal biomass production by favoring regulatory pathways that confer optimal levels of gene expression in a given environment. In this line, Tagkopoulos et al. (17) used Pearson correlations between the abundance of cell resources and the response of gene expression as a fitness function to computationally evolve the biochemical network of *E. coli* in variable environments. In this work, we propose to use the similarity of the expression profile of a

Author contributions: J.C., S.F.E., and A.J. designed research; J.C., S.F.E., and A.J. performed research; J.C., S.F.E., and A.J. analyzed data; and J.C., S.F.E., and A.J. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]To whom correspondence should be addressed. E-mail: Alfonso.Jaramillo@issb.genopole.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1200030109/-/DCSupplemental.

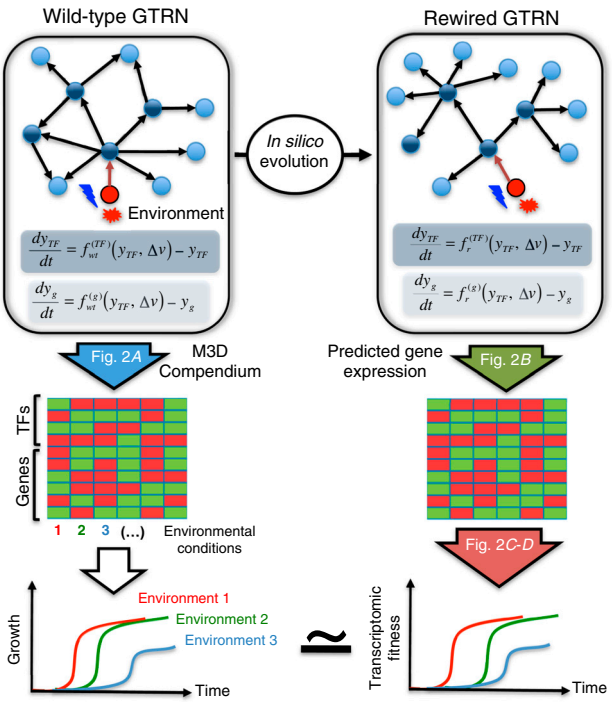

Fig. 1. Our approach for the computational evolution of a GTRN. Each step of our methodology (blue, green, and red arrows) was validated in Fig. 2.

GTRN and the wild-type (WT) as fitness function. Therefore, if we evolve a GTRN by only rewiring the transcription regulation yet keeping the same expression profile, we would expect that the solutions still have optimal growth.

Here, we analyze the transcriptional complexity required for robust growth under changing environments by developing a mathematical framework to evolve GTRNs (Fig. 1). We start by summarizing the proposed methodology for the computational evolution of GTRNs. Afterwards, we choose an organism, *E. coli*, for which an ODE for its TRN is known, and we analyze its predictability once we construct the GTRN. Next, we show that it is accurate enough to make predictions even if its topology is locally modified. Afterwards, we will analyze the resulting TRN after computational evolution under changing levels of oxygen, carbon, and nitrogen. Finally, we discuss the implications of our rewired TRN on the design principles of regulatory networks. We conclude that our methodology for rewiring genomic TRN is a useful tool to explore the design principles of transcription regulation and signaling. Our methodology will also be useful for the future re-engineering of genomes.

Computational Methods

We need to have a suitable GTRN, which we construct here by using a genome-wide model of *E. coli* gene transcription in response to selected external signals able to predict changes in cell growth after transcriptional modifications (*Materials and Methods*). The model is used to estimate kinetic parameters from experimental steady-state data (18). Given a GTRN described by a set of ODE for the concentrations of each gene product in a given genome, we propose to evolve it by an iterative procedure involving cycles of generalized mutations and selection. As generalized mutations, we consider modifications in the ODEs that could implement the move of a gene to a different operon or the addition of synthetic promoters (Fig. S1). For the selection step, we use as fitness function the similarity to a WT transcriptional profile, providing in this way the variation of cell growth. The fitness function is used in a Monte Carlo procedure to select or discard the suggested mutations (*Materials and Methods*).

Results

Environmental Adaptation of the WT GTRN. To construct the GTRN, we extended our ODE model for the TRN of *E. coli* (7) to sense environmental changes at the molecular level. We evaluated the model by quantifying how the expression of a given transcription factor (TF) changes upon the perturbation of a specific uptake factor(s) (Fig. 1 and Dataset S1). Next, we investigated how the model responds to environmental changes. We evaluated a distance, S_{exp}, between the optimal expression profile (defined as the expression profile measured for *E. coli* growing at the maximum rate for a given environmental condition) and the expression profile of the model in each environment. As it is not clear which genes will be most relevant to cell growth during our evolution, we explored six sets of genes to define S_{exp} (physiological adaptation genes, defense pathway genes, a combination of genes related to these two functions, genes that protect against abiotic stresses, genes encoding central metabolism enzymes, and all genes). Fig. S2*A* shows the optimality degree, defined as the relative growth that *E. coli* exhibits in environments that are optimal except in the concentration of a single component, such as oxygen or glucose (*Materials and Methods*) (19). Fig. S2*B* shows calculations of S_{exp} based on our model from the expression profiles predicted under 100 different environmental conditions. The largest variations of the expression score and optimality degree were obtained when selecting a gene set related to defense functions, and the smallest variation was obtained after considering genes related to enzymatic activity. This difference is expected, because the defense responses are highly inducible and specific to given environmental stimuli, whereas metabolism is able to buffer external stimulus through a critical set of metabolic pathways.

Predictability of GTRN upon Genetic and Environmental Changes. We sought to determine whether a GTRN model able to assign parameters to promoters and TF sequences predict the transcriptome of *E. coli* under different environmental conditions and/or after genetic modifications. To test our inferred model, we perform a K-fold cross-validation to ensure that gene expression profiles predicted from experimental measures of TF expression do not depend on the selection of the testing set (Fig. 2*A* and Fig. S3). We also evaluated the performance of the GTRN in predicting responses to environmental stresses and genetic changes by introducing such modifications in the model (Dataset S1). For illustrative purposes, Fig. 2*B* shows the predicted versus experimental profiles for two examples of master regulator knockouts (*fnr* and *soxS*) under aerobic and anaerobic conditions and for two environmental perturbations in which glucose, oxygen, and glycerol sources were changed. To validate S_{exp}, we compared the predicted fitness values to data from *E. coli* experimental evolution. Recently, Conrad et al. (20) characterized all acquired adaptive mutations of *E. coli* strains from a short-term laboratory evolution in minimal lactate medium. Fig. 2*C* shows a significant correlation (Pearson $r = 0.82$, 6 df, $p < 0.05$) between observed and predicted fitnesses when considering only TFs were considered in the computation of S_{exp}, thus validating our choice of the fitness function (Fig. S4*E*). Furthermore, we also attempted to predict the phenotypic response of *E. coli* after adding new regulations in its TRN (3). Fig. 2*D* show a significant correlation ($r = 0.65$, $p < 0.0001$) between growth rate and predicted fitness when only the contributions of TFs to S_{exp} was considered, corroborating that our fitness function is able to capture large changes in the TRN (*SI Materials and Methods*).

Rewiring the *E. coli* GTRN by Computational Evolution. In addition to fitness expressed as growth, S_{exp}, we needed another objective function that is related to the expected GTRN arrangement, S_{mod} (*Materials and Methods*). Fig. 3*A* illustrates the trajectories of the S_{exp} and S_{mod} functions and their weighted sums, which

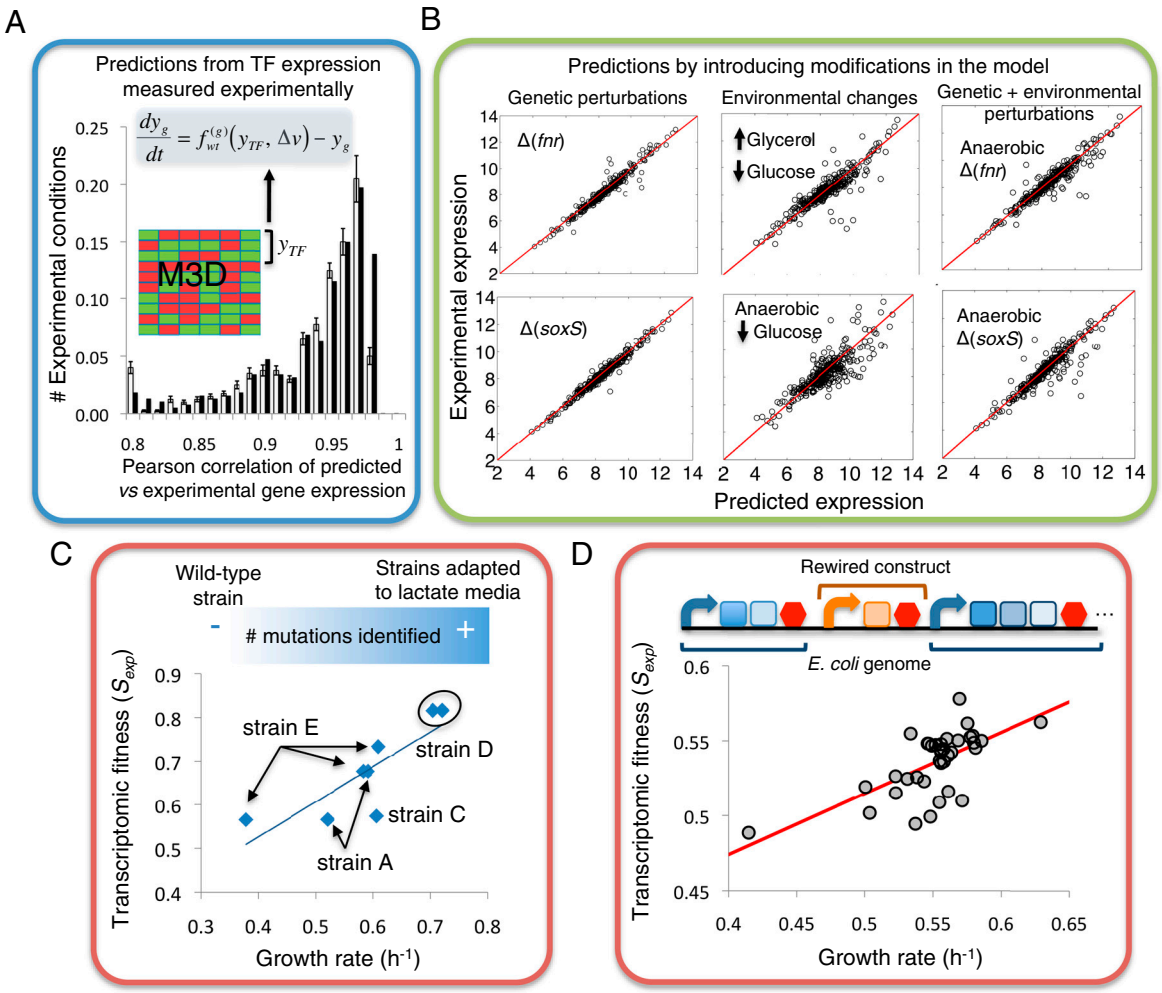

Fig. 2. (*A*) Histogram of Pearson correlations among predicted and experimental gene expressions in the 380 experimental conditions of M3D by using models trained obtained from different subsets (10-fold cross-validation) or all set of conditions (white and black bars, respectively) (Fig. S3). (*B*) Prediction of expression profiles of *E. coli* upon genetic changes (knockout of *fnr* and *soxS*), environmental perturbations (modification of oxygen and carbon availability), or both (*fnr* and *soxS* knockout under anaerobic conditions). Each dot in the scatter plots represents a value obtained from a different hybridization experiment plotted against the algorithm prediction. The red line represents the exact prediction. (*C*) Correlation between predicted fitness considering only TFs (Fig. S4*E*) and the growth rates of four strains and their intermediaries evolved in the laboratory under minimal lactate media. The GTRN of such strains was modeled by optimizing the unknown expression parameters for the mutated genes (*SI Text*). The fitness values (S_{exp}) of the WT GTRN under the different environments selected are shown in Fig. S4 (*SI Text*). (*D*) Correlation between predicted fitness considering only TFs and the growth rate of 37 strains with a rewired TRN.

defines the fitness function to be used during the in silico evolution (*SI Materials and Methods*). First, this was done for different environments by maintaining the optimal gene expression levels only for metabolic enzymes (Fig. S4 *A* and *B*). The fitness function achieved similar values during the last steps of the evolution process for all simulated replicates of the rewired GTRNs. Interestingly, we observed a significant reduction in the complexity of the rewired TRN with respect to the WT. We computed the ratio between the number of regulatory interactions ($\Xi < 0.31$, $p < 0.001$) and the number of operons ($\Theta < 0.27$, $p < 0.001$) for the rewired and WT GTRNs, which do not appear to depend on the environment. These GTRNs were optimized under the imposed constrain that only central metabolism enzymes expression must remain close to the optimal level. How does the reduction of TRN complexity depend on the selection of critical genes involved in the fitness function? To address this question (Figs. S5 and S6), we also explored the possibility that limiting the expression of only those genes related to defense and adaptation would allow larger reductions in complexity (Fig. S5*D*; $\Xi < 0.25$, $p < 0.001$; $\Theta < 0.23$, $p < 0.001$). The smallest reductions in complexity were obtained when the entire genome was restricted (Fig. S5*J*; $\Xi < 0.38$, $p < 0.001$; $\Theta < 0.33$, $p < 0.001$). Thus, high reductions in TRN complexity were obtained independently of the set of genes selected as critical predictors of transcriptomic fitness.

Next, we investigated whether genes with high functional similarity were grouped into the same operons or network modules; for example, we computed the functional similarity of all operons containing more than one gene in the rewired and random operon-organization GTRNs. Fig. 3*B* shows the highly statistically significant functional similarity of genes rewired into the same operon with respect to random evolutions (Kolmogorov-Smirnov test, $p < 0.001$; Mann-Whitney test, $p < 0.001$). It is especially interesting that the rewired GTRNs were characterized by operons containing genes of similar functions, a property that was not imposed during the evolutionary process. Specifically, the number of rewired operons with degrees of functional similarity <0.8 considerably exceeded the number of those with random organization.

Analysis of Biochemical Adaptation in Rewired GTRNs. Many signaling systems can adapt their expression programs in response to novel stimuli. Fig. S2 shows that a single, strong environmental perturbation induced WT TRN to reduce cell fitness to a minimal, but stable, value. This motivated us to investigate whether rewired systems acquired the ability to adapt to environmental changes more

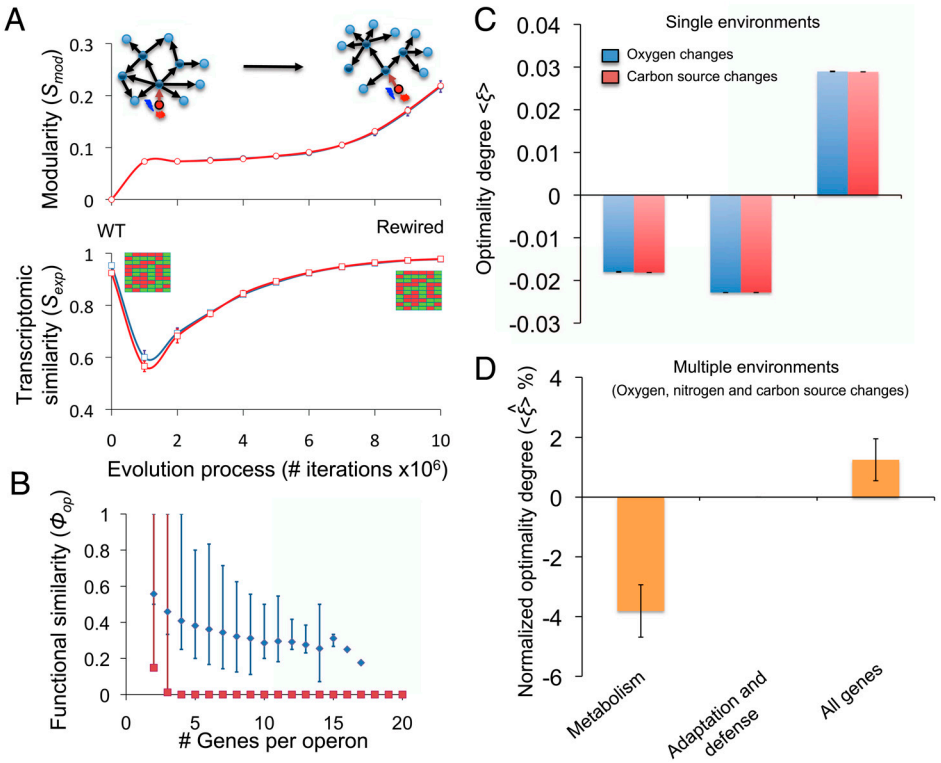

Fig. 3. (*A*) Modularity score and transcriptomic similarity for rewired GTRNs in the evolutionary steps under permissive and challenging environments (blue and red lines, respectively). Random optimizations produced significantly lower biobjective function values than those of the WT GTRN. Error bars represent standard deviations of scores obtained from 10 evolutionary processes. (*B*) Functional similarity, depending on operon size, of rewired and random GTRNs (blue and red points, respectively) that have evolved in a neutral environment. Error bars show the minimum and maximum value of functional similarity of all operons with a given size. Note that selective pressure, S_{exp}, was computed scoring only genes relating to central metabolism (see Fig. S5 for selective pressures based on stress genes or all genome). (*C* and *D*) Optimality degree ($\langle\xi\rangle$ and $\langle\hat{\xi}\rangle$, respectively) as a measurement of adaptive behavior of rewired *E. coli* GTRNs evolved under selective pressures affecting either genes coding for enzymatic activity, genes related to adaptation and defense functions, or to the entire genome. The behavior of the designed GTRNs was measured by applying single environmental perturbations (*C*) that modified external fluxes of oxygen or carbon sources (blue and red bars, respectively) or simultaneous changes in the oxygen, carbon, and nitrogen sources (*D*). Adaptation was predicted by using cell fitness constraints depending on the critical genes selected in the in silico evolution process. Error bars represent standard deviations of the scores obtained from 10 evolutionary processes.

quickly than WT systems. Fig. 4 *A* and *C* shows an example of rewired GTRN showing the operons that contain rearranged TFs regulated by three environmental factors (EFs). Interestingly, those rewired operons are controlled by a set of new regulations that highly differ from the WT operons (Fig. 4 *A* and *C*), yet maintain the original transcriptomic behavior. We explored single environmental perturbations by simulating two sets of environments. We then used the optimality degree to assess the adaptation of rewired GTRN to the environments, considering three types of selection pressure in the expression score: selecting only genes coding for enzymes involved in central metabolism, stress-related genes, or the entire genome (Fig. 3*C*). Using the first two criteria for the evolutionary process, the average of the optimality degrees $\langle\xi\rangle$ around the set of environmental perturbations was negative (i.e., cell fitness exceeded the optimal value for all re-engineered GTRNs ($\langle\xi\rangle = -0.018$ and -0.023, respectively). On the contrary, GTRNs rewired based on the third criterion achieved positive optimality degrees ($\langle\xi\rangle = 0.029$). Defining the fragility of a GTRN as its optimality degree in different environments, rewired GTRNs were more fragile; anticipatory behavior disappeared ($\langle\xi\rangle > 0.467$, 0, and 0.025 for the three evolutionary criteria mentioned, respectively) when cell fitness was computed using an expression score from a set of critical genes different from those used during the design phase. It should be noted that the optimality degree under single perturbations did not significantly depend on alterations in metabolic uptake factors.

Next, we studied systems that were re-engineered under simultaneous multiple perturbations (Fig. 3*D*). We predicted GTRN optimality by altering oxygen and carbon source uptake factors in the same range defined by single perturbations, and we added a third sensing component related to the nitrogen source by adding nitrate to the environment. As before, rewired GTRNs achieved negative or zero degrees of normalized optimality with the two first evolutionary criteria ($\langle\hat{\xi}\rangle = -3.81\%$), but for the third criterion, the average normalized optimality ($\langle\hat{\xi}\rangle = 1.25\%$) indicated that new systems retained the fitness of the optimal system.

Discussion

Design Principles of Genomic Adaptation to Environmental Changes. One important implication of our results is inference of some genome design principles (2, 4). In particular, we studied the rewired TRNs that had achieved over-optimality or lost optimality. Our rewired GTRN were more susceptible to environmental perturbations when optimality was computed using transcriptomic fitness based on a different set of genes than those selected for the computational evolution. Recent work has shown that biochemical networks have evolved to capture the multidimensional structure of diverse environments and thus form internal representations (through regulatory networks) that allow the prediction of environmental changes. For example, Tagkopoulos et al. (17) provided evidence of anticipatory behavior of *E. coli* to changes in temperature and oxygen levels that occurred over evolutionary time scales (21, 22). We examined the anticipatory ability of our rewired GTRNs by computing their optimality using transcriptomic fitness with the same set of genes used in the in silico evolutionary process. Interestingly, we found that rewired GTRNs achieved greater optimality degrees than those of WT GTRNs for both single and multiple environmental perturba-

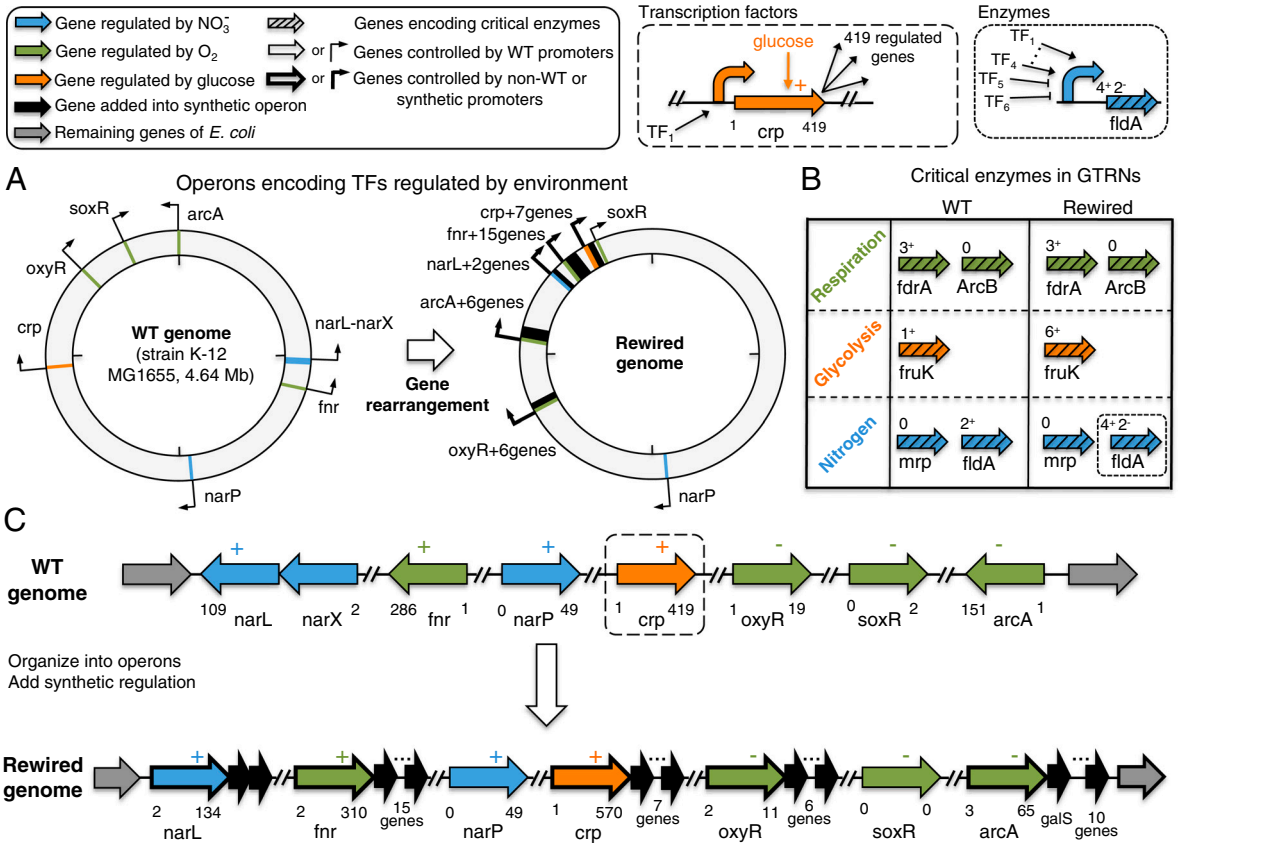

Fig. 4. Examples of the process of gene rearrangement (*A*) and rewiring (*B* and *C*) of a GTRN. Only WT or synthetic operons regulated by EFs are plotted. (*A*) WT and rewired *E. coli* genome maps showing operons regulated by the EFs [nitrogen (blue) and carbon (orange) sources, and oxygen (green)]. (*B*) Number of TFs regulating the promoters (in the WT or rewired GTRNs) of critical enzymes with rate limiting in the cellular respiration, glycolysis pathway, and nitrogen metabolism. (*C*) Operons containing TFs affected by some EFs in the WT (*Top*) and rewired (*Lower*; see also Fig. S7*A*) GTRNs. The dashed boxes surrounding genes exemplify the notation.

tions. This suggests that natural selection may be "shortsighted" (i.e., it does not anticipate large changes over the long term) and that actual TRNs have thus evolved for optimal responses to regimes of small fluctuations.

We could induct some design principles by analyzing the genes and operons involved with the uptake of oxygen, nitrate, and glucose in the WT and rewired GTRNs (Fig. 4*C*, *Upper* and *Lower*). It is insightful to recall Savageau's Demand theory (23), which in our context states that enzymes catabolizing the chemical inducers present more (less) often should be regulated by a positive (negative) transcription regulation mode. This rule is thought to provide robustness under mutational drift (5). We can check in Fig. 4*B* that the WT GTRN follows this rule by inspecting the regulation mode of the rate-limiting enzymes (genes in green boxes) associated to our chemical inducers (in high demand). As our methodology does not consider evolution in the context of a population, we do not expect it could comply with the propositions of such theory. Surprisingly, we can still see that the rewired GTRN also follows Savageau's rules for the corresponding key enzymes. Probably this is due to our choice of fitness function that forces the rewired GTRN to have similar gene expression profiles than the WT under changing environments. In addition, as our GTRN evolution only relocates genes, it is difficult to change the mode of regulation in a single evolution step. We should notice that regulatory circuits designed in synthetic biology often lack robustness to mutational drift (24). Sometimes this lack of evolutionary robustness may be desired for biosafety reasons. Further work could consider incorporating such robustness into our fitness function.

Implications of Rewired GTRNs to Genome Organization. Our results demonstrate that it is possible to rewire the GTRN of *E. coli*, achieving up to 69% reduction in the number of regulatory interactions and a 73% reduction in the number of operons, while maintaining its ability to physiologically respond to environmental perturbations. One limitation of the rewired GTRNs evolved under a single constant environment is that they will not behave like the WT GTRN under alternative environments. Fig. 4*C*, *Lower*, illustrates this point, where the *arcA* synthetic operon also contains the *galS* enzyme involved in the galactose metabolism. Fig. S7 shows how the transcriptomic fitness under galactose variation differs from the WT. In addition, we found that the rewired GTRNs contain operons that encompass several genes with similar functionality. This is an important result, given that the fitness function imposed to evaluate GTRNs performance did not consider gene function. This agrees with the experimental observation that genes within an operon have similar functions (25). Moreover, these GTRNs acquired the ability to adapt more rapidly to environmental changes, probably as a direct consequence of the reduced number of regulatory elements. Our methodology could also be applied to the de novo genome design problem if we had a perfect model of the WT GTRN. As the WT GTRN is still poorly known to aim for a faithful biological matching, it is not yet reasonable to seek any biological implementation of the rewired GTRNs, even if it would not be hard to assign a genomic nucleotide sequence for it. The de novo design of cells with synthetic genomes that are viable in a well-defined environment might require only the constitutive expression of the minimal set of genes required for life (26), but design of genomes adapted to various environments requires incorporating computational

methodologies evolving GTRNs. We expect that the improvement of GTRNs and the rapid development of technologies allowing the synthesis of novel genomes and their introduction into hosts (27–29) will allow the construction of simplified genomes.

Materials and Methods

Mathematical Genome-Scale Model. We used transcriptomic data to infer a continuous model for the transcription of all *E. coli* genes, which we then used to assign appropriate parameters to promoter and TF coding sequences. By assuming that these parameters do not depend on genomic context in most cases, we proposed our first methodology for the automatic evolution of rewired GTRNs under changing environments. Specifically, we constructed a GTRN for the WT genome that was able to predict gene regulation at the transcriptional and environmental levels (*SI Text*). For this, we adopted a linear model based on differential equations describing the time dynamics of each mRNA (7, 12) to infer kinetic parameters for promoter and TF sequences. Thus, the mRNA dynamics from the *i*th gene, y_i, is given by $dy_i/dt = a_i + \Sigma_j \beta_{ij} y_j + \Sigma_k \gamma_{ik} \Delta v_k - \delta_i y_i$, where α_i represents its constitutive transcription rate, β_{ij} represents the regulatory effect that gene *j* has on gene *i*, γ_{ik} represents the effect that environmental factor (EF),that is, the metabolic uptake factor *k*, has on the expression of gene *i*; $\Delta v_k = (v_k - v_k^{\text{opt}})$ is the difference between the uptake factor measured under a given environmental condition, v_k, and the uptake factor measured in the optimal environmental condition, v_k^{opt}; and δ_i represents the degradation and dilution rate constant.

Computational Evolution of GTRNs. The main variables required for automatic evolution of GTRNs are the same as those required for any evolutionary algorithm: (*i*) an initial GTRN, (*ii*) evolutionary steps represented by changes in the genome (Fig. S1), and (*iii*) a fitness function that evaluates the performance of each mutant GTRN (*SI Text*). For the first step, we used the GTRN of the model bacterium *E. coli*. The second step was achieved by dissecting the bacterial GTRN into elementary modules (transcriptional model of the *E. coli* WT GTRN, http://repository.issb.genopole.fr/frontal/Technology/Tools/Carrera_SupDat2.xml/at_download/file), to which evolutionary rules were applied.

One design approach that we used involved the computational evolution of the GTRN, where we pursued two goals simultaneously (*SI Materials and Methods*): (*i*) simplifying the internal structure of the *E. coli* GTRN, and (*ii*) maintaining the external system function. To maximize the modularity of the system and thus simplify the TRN, we defined a measure based on the entropy of the TRN, $S_{\text{mod}} = 1 - \sum_{op}^{N_{op}} k_{op} \log_{N_g} k_{op}^{-1}$, where $k_{op} = N_g^{op}/N_g$. N_g^{op} represents the number of genes in the operon *op*, N_g is the number of nonconstitutive genes in the WT GTRN, and N_{op} is the updated number of operons contained in the rewired GTRN. We also aimed to maximize the similarity of the expression profiles of the WT (y^{opt}) and rewired (y^{env}) GTRN for a set of extreme environments (N_{env}) and for a set of critical genes that guarantee the functionality of the rewired GTRN, $S_{\text{exp}} = [\prod_{\text{env}} \rho(y_g^{\text{opt}}, y_g^{\text{env}})]^{\frac{1}{N_{\text{env}}}}$, where *g* denotes genes included in a set of critical genes that guarantee the optimal growth of the cell. We used the TRN model integrated with signal transduction to measure that similarity. Considering these two aims, we developed an optimization algorithm based on the mutation rules described in Fig. S2 to rewire the WT *E. coli* GTRN (*SI Materials and Methods*). Genes that are controlled by constitutive promoters were not involved in the computational evolution. These genes could always be regrouped in a straightforward way by assuming that they could be collapsed into large operons regulated by a gradient of different expression levels.

GTRN Optimality Degree. We assumed that cell fitness could be estimated in terms of the S_{exp} objective function. This allowed the study of GTRN adaptation under changing environments in one ($\Delta v_{k=i} \neq 0$ and $\Delta v_{k\neq i} = 0$) or multiple ($\Delta v_k \neq 0 \forall k$) directions (14). To do this, we defined the optimality degree, $\xi_{\Delta v_k^*}$, in a target environment characterized by Δv_k^* and different from the optimal environment as the difference between S_{exp} evaluated in an environment containing $\Delta v_k = 0$ (i.e., fitness in the optimal condition) and that evaluated in the target environment containing Δv_k^*. Hence, we distinguished between positive and negative error adaptation corresponding to environmental states where cell fitness achieved sub- or over-optimal growth, respectively.

Functional Analysis of GTRNs. Genes contained in the operons of all rewired GTRNs were functionally identified using 184 biological functions in GO (30). We defined the degree of functional similarity (ϕ_{op}) of a given operon, *op*, as the ratio between the maximum number of genes with the same functionality and the operon size. We imposed $\phi_{op} = 0$ for those operons containing only one gene because more than one gene was needed to assess functional similarity.

ACKNOWLEDGMENTS. This work was supported by FP7-ICT-043338 (Bacterial Computing with Engineered Populations), ATIGE-Genopole, TIN2006-12860 (Ministry of Science and Innovation [MICINN]), and the Fondation pour la Recherche Medicale grants (to A.J.). S.F.E. is supported by grant BFU2009-06993 (MICINN). We thank B. Palsson, T. Conrad, and M. Isalan for providing us with experimental data from their recent publications, J. Forment for help with computer resources; R. Estrela, G. Rodrigo, for discussions; J. Sardanyés, T. Landrain, L. Janniere, I. Junier, M. P. Zwart, and F. Kepes for critical reading of the manuscript; and the comments provided by anonymous reviewers.

1. Khalil AS, Collins JJ (2010) Synthetic biology: Applications come of age. *Nat Rev Genet* 11:367–379.
2. Bhardwaj N, Kim PM, Gerstein MB (2010) Rewiring of transcriptional regulatory networks: Hierarchy, rather than connectivity, better reflects the importance of regulators. *Sci Signaling* 3:ra79.
3. Isalan M, et al. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452:840–845.
4. Bashor CJ, et al. (2010) Rewiring cells: Synthetic Biology as a tool to interrogate the organizational principles of living sytems. *Annu Rev Biophys* 39:515–537.
5. Alon U (2007) *An introduction to systems biology: Design principles of biological systems* (Chapman & Hall/CRC, London).
6. Ronen M, Rosenberg R, Shraiman BI, Alon U (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci USA* 99:10555–10560.
7. Carrera J, Rodrigo G, Jaramillo A (2009) Model-based redesign of global transcription regulation. *Nucleic Acids Res* 37:e38.
8. Carrera J, Rodrigo G, Jaramillo A (2009) Towards the automated engineering of a synthetic genome. *Mol Biosyst* 5:733–743.
9. Chan LY, Kosuri S, Endy D (2005) Refactoring bacteriophage T7. *Mol Syst Biol* 1:2005.0018.
10. Bonneau R (2007) A predictive model for transcriptional control of physiology in a free living cell. *Cell* 131:1354–1365.
11. Covert MW, et al. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:92–96.
12. Gardner TS, et al. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* 301:102–105.
13. Carrera J, Rodrigo G, Jaramillo A, Elena SF (2009) Reverse-engineering the *Arabidopsis thaliana* transcriptional network under changing environmental conditions. *Genome Biol* 10:R96.
14. Francois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci USA* 101:580–585.
15. Rodrigo G, Carrera J, Jaramillo A (2011) Computational design of synthetic regulatory networks from a genetic library to characterize the designability of dynamical behaviors. *Nucleic Acids Res* 39:e138.
16. Airoldi EM, et al. (2009) Predicting cellular growth from gene expression signatures. *PLoS Comput Biol* 5:e1000257.
17. Tagkopoulos I, Liu YC, Tavazoie S (2008) Predictive behavior within microbial genetic networks. *Science* 320:1313–1317.
18. Faith JJ, et al. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 5:e8.
19. Ma W, et al. (2009) Defining network topologies that can achieve biochemical adaptation. *Cell* 138:760–773.
20. Conrad TM, et al. (2009) Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol* 9:R118.
21. Perkins TJ, Swain PS (2009) Strategies for cellular decision-making. *Mol Syst Biol* 5:326.
22. Koide T, Pang WL, Baliga NS (2009) The role of predictive modeling in rationally re-engineering biological systems. *Nat Rev Microbiol* 7:297–305.
23. Savageau MA (1998) Demand theory of gene regulation. I. Quantitative development of the theory. *Genetics* 149:1665–1676.
24. Sleight SC, Bartley BA, Lieviant JA, Sauro HM (2010) Designing and engineering evolutionary robust genetic circuits. *J Biol Eng* 4:12.
25. Keseler IM, et al. (2011) EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res* 39 D583–D590.
26. Forster AC, Church GM (2006) Towards synthesis of a minimal cell. *Mol Syst Biol* 2:45.
27. Lartigue C, et al. (2007) Genome transplantation in bacteria: Changing one species to another. *Science* 317:632–638.
28. Dymond JS, et al. (2011) Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* 477:471–476.
29. Temmea K, Zhaob D, Voigt CA (2012) Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. *Proc Natl Acad Sci USA* 109:7085–7090.
30. Ashburner M, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29.

Supporting Information

Carrera et al. 10.1073/pnas.1200030109

SI Text

SI Materials and Methods. ***Construction of a genome-wide transcriptional model for gene expression.*** The model contains 4298 nonredundant genes, 330 of which are putative transcription factors (TFs) (1). Such a model allows the assignment of mathematical parameters to promoters and TF sequences, which we have assumed to be independent of genomic context. Recent studies have collected data describing thousands of interactions between environmental factors (EFs) and TFs that are involved in sensing environmental perturbations. These interactions were coupled to the transcriptional regulatory network (TRN) model such that uptake factors modified the predicted expression of several TFs.

Time was conveniently scaled such that $\delta_i = 1$ and the model was assumed to be in steady-state $y_i = \hat{\alpha}_i + \sum_j \beta_{ij} y_j$, where $\hat{\alpha}_i = \alpha_i + \varepsilon_i + \sum_k \gamma_{ik} \Delta v_k$, because fitting the appropriate messenger RNA (mRNA) degradation constant would require time series data (2). To calibrate TF expression, the newly redefined constitutive transcription rate included a perturbative term (ε_i) that fit only the TF expression profile (y^{opt}) for the defined optimal condition $\varepsilon_i = \sum_j (1 - \beta_{ij}) y_j^{\text{opt}} - \alpha_i$. Each TF expression is bounded $\varphi y_i^{\min} \le y_i \le \varphi^{-1} y_i^{\max}$ by a range interval defined by the minimum ($y_i^{\min}$) and maximum ($y_i^{\max}$) value of all experimental measurements for that TF in the microarray compendium (3); $\varphi \ge 1$ is a tunable parameter that decreases the gene expression range to improve the predictive capacity of the presented model under environmental and genetic perturbations.

To construct the genomic transcriptional regulatory network (GTRN) model, we used steady-state mRNA expression profiles derived from transcriptional perturbations collected in *M3D* version 4.5 online (3). We identified 330 TFs by searching for the key phrase "transcription factor" in the functionally annotated *E. coli* genome from RegulonDB (version 5) (4). The dataset contains preprocessed expression data from 380 hybridization experiments using 4,289 probe sets spotted on an Affymetrix GeneChip. Data were normalized using the robust multiarray average method (5) and represented on the $\log_2$ scale. The inference procedure consisted of three nested steps. In the first step, global network connectivity was inferred using the InferGene algorithm (1). This method uses mutual information (MI) with local significance (z-score computation) to compute the number of transcriptional regulations in the genome (6). Hence, each potential interaction between a regulator and a gene receives a z-score, which provides an estimate of MI. This approach eliminates some false correlations and indirect influences (6). Subsequently, we selected a z-score threshold for cutoff. We included transcriptional regulations that were experimentally compiled in RegulonDB (4), but not those inferred by our procedure. Then, multiple regressions based on ODEs were performed to estimate the kinetic parameters of the regulatory model. The resulting file containing the model of the *Escherichia coli* (*E. coli*) transcriptional response to its environment is available as transcriptional model of the *E. coli* WT GTRN, http://repository.issb.genopole.fr/frontal/Technology/Tools/Carrera_SupDat2.xml/at_download/file.

The wild-type (WT) transcriptional network contains 2,987 inferred regulatory interactions with z-scores over the selected threshold of 5. The network also contains 3,388 interactions from the reference regulatory set constructed based on RegulonDB (4); 179 of these experimental interactions also belonged to the inferred test. The performance of the inferred TRN model topology was evaluated using a reference network defined by genes with known transcriptional regulation. Only interactions among genes included in this reference set were considered. The fraction of interactions that were correctly predicted by the model (the precision, P) and the fraction of all known interactions that were discovered by the model (the sensitivity, S) were used to compute a global performance statistic defined as $F = 2PS/(P + S)$ (7). This TRN has a global performance of $F = 11.8\%$ (35.1% precision and 7.1% sensitivity) in predicting the regulations identified in RegulonDB. While this provides far from complete understanding of the regulation network of *E. coli*, the model constructed demonstrates sufficient predictive power to be used as starting point for our evolution (Dataset S1).

Construction of a transcriptional regulatory network that integrates signal transduction. Biological systems optimize their regulation by monitoring changes in their environment. Gene expression is largely controlled at the level of transcription by TFs. In addition to a DNA-binding domain, TFs often have structural domains that can bind specific metabolites. Thus, we increased the TRN complexity by including 299 external metabolic fluxes (8) as environmental factors (EFs). These EFs are direct links from the environment to the genetic network, affecting the expression of several TFs, and are common signals for endogenous and exogenous changes in cell state.

To link the environment to the regulatory network of the genome, we used two sets of experimentally obtained EF-TF interaction data published by Martínez-Antonio et al. (9) and Wall et al. (10). However, only regulations in which the EF represents one of the 299 external metabolic fluxes defined in the work of Feist et al. (8) were considered, reducing the set to 65 interactions (EF-TF) involving 50 EFs and 53 TFs. The transcriptional sensing system that was added to the TRN incorporated three types of sensors: (*i*) 14 transported metabolites (E-TM) that are sensed externally, (*ii*) four TFs that sense metabolites that are generated internally (I-SM), and (*iii*) 37 TFs that sense metabolites that are both transported and generated in the cytoplasm, that is, a hybrid system (H) (9). Hence, we focused our study on one-component signal transduction pathways because these are more widely overrepresented in bacteria and display a greater diversity of domains than do two-component systems (11).

We computed γ_{ik} as a perturbation of the expression in the optimal condition of the gene i due to an environmental change that also perturbs the optimal state of the metabolic flux k, $\gamma_{ik} = \frac{\vartheta y_i^{\text{opt}}}{v_k^0 - v_k^{\text{opt}}}$, where ϑ is a parameter that represents the normalized variation of the optimal expression. This parameter is optimized to fit the experimental gene expression under genetic and environmental perturbations (File S3). If j or k have no effect on the expression of i, then $\beta_{ij} = 0$ and $\gamma_{ik} = 0$; in fact, only regulatory effects of EFs on TFs are considered. We have not incorporated the effects of cooperation in transcription regulation. We have used public microarray hybridization data (3) from an Affymetrix chip normalized using RMA (5). This microarray compendium contains data from 380 experiments.

Two parameters were optimized: $\varphi = 0.9$ defines the model gene expression range, and $\vartheta = 0.5$ characterizes the variation in the WT expression of a given TF due to the influence of a specified external metabolic flux. These parameters were optimized to fit several predicted gene expression profiles from 31 experiments (contained in the M3D compendium [3]) corresponding to transcriptional and environmental perturbations (File S3). Specifically, we used data from 16 knockouts of transcriptional master regulators (*appY*, *arcA*, *fnr*, *soxR*, *soxS*, *recA*, *fis*, *yncC*, and *rpoS*), eight environmental perturbations of oxygen and carbon sources

(glucose, acetate, glycerol, and proline), and seven conditions combining both types of perturbations (12).

Selecting environments to generate different degrees of optimality in the WT GTRN. Five sets of six environments defined by external oxygen flux, carbon source (external glucose flux), and nitrogen source (external NO_3^-) were selected based on the decrease that each caused in the expression score, S_{exp}. Specifically, we included environments in each set based on five levels of decreases in S_{exp} that range from 0 to 10%. All sets include the environment associated with the optimal condition, creating different ranges of environmental variability for each set. Fig. S4 *A* and *B* shows the expression score multiplied by the expression score in the optimal environment under conditions whose distance to the optimum ranges between 0 and 2% (weak perturbations, Fig. S4 *A* and *B*) and between 8 and 10% (strong perturbations, Fig. S4 *C* and *D*). The expression scores in *A* and *C* were computed considering only genes coding for enzymes, and those in *B* and *D* were computed considering only genes related to adaptation or defense.

Genome-wide optimization procedure. Our algorithm searches possible reconfigurations of the global transcriptional regulation of *E. coli* such that the resulting modular genome contains all genes in a minimal set of operons, thus decreasing the number of transcriptional regulatory elements, and with the constraint that the overall gene expression of the rewired GTRN shall be as close to the WT as possible. We used Monte Carlo simulated annealing (13) to perform the optimization in the space of all possible rewired transcriptional networks. The size of this combinatorial space is governed by the previously characterized variability in the *E. coli* natural promoters, and the diversity of synthetic promoters was obtained during the optimization process. As the starting condition, we assumed that the expression of each gene was controlled only by its natural promoter. Based on the transcriptional regulation landscape size, we defined two sets of optimization processes. In the first set, we introduced small transcriptional modifications in the GTRN at each step of the optimization by either changing the regulation of a gene (moving it downstream of another promoter) or eliminating regulation by natural or synthetic promoters according to the following rules (Fig. S1):

i. Move gene g belonging to operon op and regulated by a nonconstitutive promoter $P(op)$ to another operon op' regulated by a different nonconstitutive promoter $P'(op')$. When g moves to op', we add all regulatory operators of its natural promoter to P' (Fig. S1*A*). However, the fact that g leaves P implies that if the gene is regulated by a promoter different from its natural promoter, then P will lose all inserted operators due to the regulatory effect of P on g (Fig. S1*B*). Coexpression of all genes expressed from a given operon was imposed.
ii. Remove an operator from a synthetic promoter (Fig. S1*C*). Only operators associated with TFs are likely to be removed. Unlike transcriptional regulations, interactions of TFs associated with the binding with EFs remain linked to their corresponding genes throughout the optimization process.

To simplify the genome network structure and improve algorithm convergence, the probability of removing a regulation was made much larger than the probability of changing a gene's promoter (e.g., 10-fold). Expression behavior of the newly created genome and compute its new objective function (S_{new}), which depends on the full transcriptome predicted under a set of environments and the new modular organization of the operons. If the suggested mutation improves $S(S_{new} \geq S)$, then it is accepted. Otherwise, it is accepted with probability $e^{(S-S_{new})/T}$, where T is a Boltzmann temperature parameter that decreases exponentially with the number of iterations. Hereafter, we loop back and introduce a new transcriptional modification.

Objective functions for in silico evolution. After generating the predictive model for the GTRN, we attempted to automatically rewire GTRNs by implementing an in silico evolution algorithm in which a fitness function is used to select for beneficial GTRN modifications during the evolution process. We aimed to rearrange genes (refactorization) within the GTRN of *E. coli* such that the information content of the distribution of genes in operons could be increased. We hypothesized that this would produce a genome with fewer operons but retaining the entire original set of genes. Therefore, we considered a measure based on Shannon entropy (14) as the first objective function. This measure is computed from the distribution of genes in the operons as $S = 1 - \sum_{op}^{N_{op}} k_{op} \log_{N_g} k_{op}^{-1}$, where $k_{op} = N_g^{op}/N_g$. N_g^{op} represents the number of genes in the operon op, N_g is the number of nonconstitutive genes in the WT GTRN, and N_{op} is the updated number of operons contained in the rewired GTRN. Genes initially controlled by constitutive promoters were not involved in the optimization because we assumed that unregulated genes with similar basal expression levels could be grouped into operons controlled by constitutive promoters that provide similar expression levels regardless of the environment. By defining the logarithm base as N_g, we ensured that S_{mod} ranges from 0 to 1, thereby obtaining null modularity for the WT genome. We assumed in our model that the sizes of all operons in the WT GTRN are equal to one because genes that are known to be controlled in the same operon did not share the same experimental interactions with TFs collected in RegulonDB (4) or inferred by the InferGene algorithm (1). Thus, precision and recall in the inference of the GTRN were maximized. The second objective function was defined as the distance from the WT gene expression profile to the predicted profile under various environmental conditions (15). This similarity was measured as the Pearson correlation coefficient (ρ) obtained when the predicted expression profiles for a set of extreme environments (N_{env}) were compared to the WT expression, $S_{exp} = [\prod_{env} \rho(y_g^{opt}, y_g^{env})]^{\frac{1}{N_{env}}}$, where g denotes genes included in a set of critical genes that guarantee the optimal growth of the cell (e.g., genes encoding enzymatic activity). We defined three sets of critical genes: (*i*) genes coding for enzymatic activity, (*ii*) genes related to the stress response, and (*iii*) all genes. Ultimately, we defined a bi-objective function based on the weighted sum of both objectives, $S(\lambda) = \lambda S_{exp} + (1-\lambda) S_{mod}$; thus, selecting a given weighting factor, $\lambda \in [0, 1]$, the bi-objective problem relies on maximizing S by the Monte Carlo simulated annealing optimization protocol. We used $\lambda = 0.5$ for the simulations.

In silico GTRN evolution by adaptive mutation. With slight alterations, our methodology was able to predict the behavior of intermediary *E. coli* strains generated from laboratory evolution by local adaptive mutations in minimal lactate media (16). They measured growth rates and identified adaptive mutations using whole-genome sequencing for all evolved strains at specific time points. Interestingly, several mutations were identified in highly connected TFs in the TRN (*crp*, *ydcI*, and *hfq*) in a gene related to transcription termination (*rho*) and in a gene responsible for recycling RNA polymerases (*hepA*). We predicted the transcriptome for each of these strains by modifying our *E. coli* network model to introduce a different gene expression value for each mutated gene. We then determined the S_{exp} fitness function of the predicted expression profile by predicting the transcriptome of a strain with the mutated genes set at optimal transcription levels and then calculating the distance between

the mutant strain and the optimal strain evolved for adaptation in lactate culture.

We chose strains that were more adapted to the new media as the optimal model, in contrast to the in silico evolution by GTRN rewiring in which the WT strain was the optimal model. Consequently, this altered the expression profile required to maintain optimal cell behavior. Hence, by introducing all adaptive mutations to the WT GTRN model, we were able to simulate the optimal gene expression profile, y_g^{adapted}. We replaced the corresponding ODEs of the mutated genes ($\hat{g}$), $dy_{\hat{g}}/dt = \alpha_{\hat{g}} + \sum_j \beta_{\hat{g}j} y_j + \sum_k \gamma_{\hat{g}k} \Delta v_k - \delta_{\hat{g}} y_{\hat{g}}$, with constant expression values to simulate the new steady state of that mutated gene, $y_{\hat{g}} \in [\varphi y_{\hat{g}}^{\min}, \varphi^{-1} y_{\hat{g}}^{\max}]$. Note that to simulate the appropriated minimal media, we imposed $\Delta v_k = 0$ for all metabolic uptake factors excepting the lactate ($\Delta v_{\text{lactate}} = 20$) and glucose ($\Delta v_{\text{glucose}} = -10$) uptakes. Solving the new system of ODEs that incorporates the adaptive mutations,

$$dy_i/dt = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i$$

$$y_{\hat{g}} = Y_{\hat{g}}, Y \in [\varphi y_{\hat{g}}^{\min}, \varphi^{-1} y_{\hat{g}}^{\max}]$$

we simulated different gene expression profiles, $y_g = y_g(Y_{\hat{g}})$, as functions of the steady-state expression of the mutated genes, $Y_{\hat{g}}$, for intermediary adaptive *E. coli* strains. We computed transcriptomic fitness as the distance measured by ρ from the gene expression profile of the strain most adapted to the new environment with lactate (y_g^{adapted}) to the predicted profile incorporating adaptive mutations ($y_g = y_g(Y_{\hat{g}})$),

$$S_{\exp}(Y_{\hat{g}}) = \rho(y_g^{\text{adapted}}, y_g(Y_{\hat{g}})).$$

Note in Figs. 2*C* and 4*E* that we selected $S_{\exp}(Y_{\hat{g}})$ to optimize the correlation between growth rate and $S_{\exp}$ for the different intermediary steps of each strain evolved. Interestingly, we found that the gene mutations that caused maximal $S_{\exp}(Y_{\hat{g}})$ also guaranteed maximal correlations.

Note that similar correlations were observed when considering the contributions of central metabolism enzymes, genes related to stress, or the full genome with respect to the Fig. 2*C*. Overall, we showed that growth rates predicted using in silico evolution reached high correlations ($r > 0.72$, $p < 0.05$; Fig. S4*E*).

Predicting the growth rate of rewired transcriptional networks of E. coli. Our methodology was able to capture the behavior of *E. coli* strains with TRNs rewired by adding on a WT genetic background new links from different recombinations of promoters with TFs. A recent study by Isalan et al. (17) systematically explored such problem by expressing endogenous promoters controlling different TFs or σ-dependent genes and measuring the growth rate of each rewired strain. In our study, we did not consider promoter region—open reading frame fusions that were constructed on high copy number plasmids because our model is limited to predict gene expression from the bacterial genome. Therefore, we selected 38 strains from Isalan et al. (17) collection in which the rewired construct was stably integrated in the *E. coli* chromosome. For these strains, we computed their growth rate as the maximum value of $\Delta(\ln \text{OD}_{600})/\Delta t$ (with $\Delta t = 1$ h), achieving values between 0.39 h^{-1} and 0.63 h^{-1}. The strain with the construct of the TF, *rpoE*, controlled by the promoter *appY* was not included in the dataset because it showed the largest lag phase and slowest growth rate compared to the rest of the strains, indicating that the levels of gene expression are not necessarily in steady state. Therefore, this strain violated our assumption of steady-state gene expression as a proxy to fitness, $S_{\exp}$.

Hence, by introducing the modification imposed by the rewired construct to the WT GTRN model, we were able to simulate the gene expression profile of the rewired TRN and consequently predict fitness. We modified the corresponding set of ODEs that models the expression of the TF (TF_c) encoded in the construct, *c*, and controlled by the promoter, *p*. Specifically, we added the basal rate, and that determines the gene expression of the genes controlled by *p* to the ODE models TF_c. Solving the new system of ODEs,

$$dy_i/dt = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i, \ \forall i \neq \text{TF}_c$$

$$dy_i/dt = \alpha_i + \sum_j \beta_{ij} y_j + \sum_k \gamma_{ik} \Delta v_k - \delta_i y_i + \alpha_p + \sum_j \beta_{pj} y_j + \sum_k \gamma_{pk} \Delta v_k, i = \text{TF}_c$$

we simulated the gene expression profile of the rewired TRN in order to compute the fitness $S_{\exp}$. Note that to simulate the appropriated medium, we imposed $\Delta v_k = 0$ for all metabolic uptake factors excepting the glucose uptake ($\Delta v_{\text{glucose}} = 50$) to provide an excess of carbon source.

SI Experimental Procedures. In this section, we describe additional experiments done to validate or to extend our methodology for in silico evolution of GTRNs. This required the integration of current known transcriptomic and signaling data into a global model consisting of differential equations, allowing the assignment of parameters to promoter and TF coding sequences. We examine the outcome of this model construction and its corresponding properties. Next, we validated the GTRN using experimental expression profile data (see the section *Prediction of Expression Profiles upon Genetic and Environmental Changes*, *SI Text*). After a suitable model was generated, we validated the fitness function (as defined in the subsections In silico *GTRN Evolution by Adaptive Mutation* and *Predicting the Growth Rate of Rewired Transcriptional Networks of E. coli* of the *SI Methods and Materials*) to be used in our in silico evolutionary procedure. We used experimental results from a laboratory evolution experiment to show that measured growth rate differences correlate with variations in fitness. This allowed us to perform an in silico GTRN evolution simulation with the aim of rewiring the *E. coli* GTRN to simplify its internal structure by reducing the number of operons and indirectly minimizing the interactions necessary for the TRN. We found that we could dramatically reduce the number of operons while maintaining the organism's response to fluctuating environments. We also analyzed other properties of the synthetic TRN (*Topological Properties of Rewired GTRNs*, *SI Text*), such as its topology and adaptation to varying environments (*Biochemical Adaptation of the Rewired GTRNs*). Finally, we examine some design principles that can be inferred from our results in the section *Cellular Environments Selectively Correlate with Complexity of Rewired GTRNs*.

Prediction of expression profiles upon genetic and environmental changes. We compared predicted expression levels of all TFs ($\hat{y}_{gc}$) for an experimental condition, *c*, with respect to the corresponding empirical measurement, y_{gc}, using the normalized Euclidean distance (e_c) and the Pearson correlation coefficient (ρ_c) in some microarray experiments selected from *M3D* compendium.

In a first step, we used our model (trained by all conditions of the *M3D*) to predict experimental expression profiles by introducing the experimental values of TF expression ($y_{\text{TF},c}$); applying the model, $\hat{y}_{g,c} = \alpha_g + \sum_j \beta_{ij} y_{\text{TF},c}$, we predicted expression of

the rest of the genes and, consequently, we computed ρ_c for the 380 conditions of *M3D* (Fig. 2*A*, black bars). In addition, we performed a 10-fold cross-validation to predict $\hat{y}_{gc}$ by using models constructed with a training set excluding random test conditions from *M3D*. We computed ρ_c^{CV} in the test conditions (Fig. 2*B*, white bars). Interestingly, ρ_c predicted by the model trained by all experimental conditions and ρ_c^{CV} resulting of obtaining models with a lower number of conditions in the training set do not present differences statistically significant (Mann-Whitney test, $p > 0.1$). We also analyzed model accuracy to predict ρ_c^{CV}, depending on the size of the training set selected from M3D compendium. Surprisingly, models inferred with training sets containing from 0–90% of conditions, provided expression errors in the testing sets ranged between 4.2% and 4.7% (Fig. S3).

Next, we used our model (trained by the entire M3D compendium) to predict expression profiles in 31 experimental conditions (File S3) from M3D in which we found TF knockouts and environmental perturbations. To simulate these genetic changes, that is, TF knockouts, we removed the corresponding regulatory coefficients $\beta_{g,\mathrm{TF}} = 0(\forall g)$ in our model. Analogously, to simulate environmental perturbations, we modified the corresponding uptake factors, Δv_{EF}. Hence, for the set of predictions in experimental conditions with genetic changes (see examples in Fig. 2*B*, *Left Column*), we obtained values of $\rho_c > 0.80$ and $e_c < 4.32\%$, with the exception of the *yncC* knockout ($\rho_c = 0.74$) made by Faith et al. (6). In conditions in which oxygen and carbon sources were perturbed, we estimated $\rho_c > 0.74$ and $e < 7.3\%$ (see examples in Fig. 2*B*, *Central* and *Right* columns). Moreover, we performed a 10-fold cross-validation to compute the statistical significance of ρ_c and e, depending on the training set used. For that, we re-computed such scores by using models trained by sets of conditions in which 10% of random conditions from M3D were excluded. Interestingly, those models provided values of ρ_c (Mann-Whitney test, $p > 0.365$) and e (Mann-Whitney test, $p > 0.361$) extremely close to those obtained by using the model trained by all conditions. This corroborated a strong independence of the training set selected for predicting gene expression under genetic and environmental changes.

Topological properties of rewired GTRNs. The evolved configuration based on interconnected building blocks provided a significant increase in the diameter and characteristic path length of the rewired networks. Similarly, rewired GTRNs tend to lose the hierarchical scale-free system characteristics of the WT TRN (18). Whereas the slope of the log-log regression for the average clustering coefficient with the number of genes with k-connections is close to one for the TRN, it was significantly less than 1 for the rewired GTRNs. Furthermore, the power-law that fits the incoming ($\gamma_{\mathrm{incoming}}$) and outgoing ($\gamma_{\mathrm{outcoming}}$) connectivity distributions of the rewired GTRNs are both smaller than those observed for the WT TRN, corroborating the observation that in re-engineered TRN, a large number of TFs are responsible for activating different biological modules that emerged spontaneously.

Next, we analyzed the changes in promoter type across the entire genome after in silico evolution. The number of genes controlled by promoters that interact with only one TF was significantly smaller for the rewired than for the WT TRN, and the number of genes controlled by two or more TFs significantly increased (Table S1). The minimum percentage of operons controlled by a synthetic promoter in the rewired GTRNs was 17%–20%, depending on the fitness definition (i.e., whether fitness considered only genes coding for enzymes involved in central metabolism or only genes related to stress responses, respectively). Consequently, the minimum percentage of synthetic regulations added was greater than 9.5%.

We also studied the dependence between the number of TFs regulating promoters and operon size. Fig. S6 illustrates that the operon sizes of rewired GTRNs that were evolved under permissive and challenging environments were optimally controlled by a constant number of TFs, usually two or three regulators, depending on the selective pressure used in the design function. This analysis excluded large operons containing more than 20 genes, for which the number of regulatory factors exceeded six TFs.

Cellular environments selectively correlate with complexity of the rewired GTRNs. We compared the internal structures of networks evolved under permissive and challenging environments (Table S1) to determine whether the environmental conditions imposed in the evolutionary process confer any specific characteristics to the TRN of the rewired GTRNs. The clustering coefficients (*CC*s) of the rewired TRN were highly reduced with respect to the WT TRN, illustrating that the rewired networks are composed of large modules that induce additional coregulation. Interestingly, rewired GTRNs in challenging environments show higher *CC*s than those evolved in more permissive environments, a difference supported by the positive Pearson correlation observed between the *CC*s and the gradient of environmental stress ($r = 0.63$, $p < 0.01$) when S_{exp} was computed considering stress genes only.

We then analyzed the relationship between the reduction in GTRN complexity and the environment in which the networks evolved. We found no significant correlation between increased environmental challenge (measured as the variation in cell fitness) and the complexity of the rewired GTRN (measured either as the number of operons or as the number of regulatory interactions) (Table S1). Surprisingly, the positive correlation observed between *CC*s and environmental stress did not contribute to a significant relationship in terms of the complexity of the rewired TRN (SI). Next, we focused only on the rewired operons that were regulated by promoters whose TFs interacted with EFs. Surprisingly, we found a significant difference between the average size of the rewired operons in permissive and challenging environments for low operon size (LOS) (Fig. S5). Specifically, we found significant changes when the selection pressure forced optimization of all gene expression (Fig. S5 *K* and *L*) (Mann-Whitney test, $p < 0.001$) or the optimization of stress-related gene expression (Fig. S5 *E* and *F*) (Mann-Whitney test, $p < 0.0001$). In fact, for stress-related gene expression, we also observed significant changes in the average size of the rewired operons for high operon size (HOS). This is direct evidence that environments in which cells perform poorly (i.e., with very poor fitness) favored the emergence of operons containing a large number of genes coexpressed under promoters whose TFs respond to this environment.

Biochemical adaptation of the rewired GTRNs. Two sets of environments were simulated to explore single environmental perturbations: (*i*) a set of 100 random perturbations that varied oxygen availability from a fully anaerobic environment to an environment with a rate that was fourfold greater than the optimal flux value (75 mmol g^{-1} h^{-1}) and (*ii*) a set of 100 perturbations that changed the availability of glucose as the carbon source, ranging from the negative value of the optimal uptake flux to the positive value (i.e., −20 mmol g^{-1} h^{-1} to 20 mmol g^{-1} h^{-1}).

The rewired GTRNs maintained the global physiological response under both optimal and changing environments. In addition, we found that there was an increase in the complexity of the internal structure related to the signal transduction for all rewired GTRNs. More specifically, GTRNs that evolved under the most extreme environments required a greater reorganization of critical genes under promoters that could sense greater numbers of environmental interactions. Interestingly, GTRNs that were rewired under stressful environments showed higher *CC*s than those that evolved under more permissive environments. An intuitive explanation for this observation relies on the differences in the selective pressures imposed by both types of environments.

Survival and replication in a stressful environment represent stringent selection, requiring the coordinated expression of all genes involved in survival. By contrast, replicating in a permissive environment may be equated to soft selection and therefore does not require the coordination of expression because the cells remain able to exploit some components of their environment.

1. Carrera J, Rodrigo G, Jaramillo A (2009) Model-based redesign of global transcription regulation. *Nucleic Acids Res* 37:e38.
2. Bar-Joseph Z (2004) Analyzing time series gene expression data. *Bioinformatics* 20:2493–2503.
3. Faith JJ, Driscoll ME, Fusaro VA, Cosgrove EJ, Hayete B, Juhn FS, Schneider SJ, Gardner TS (2008) Many microbe microarrays database: Uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res* 36:D866–D870.
4. Salgado H, Gama-Castro S, et al. (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 34:D394.
5. Irizarray RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
6. Faith JJ, et al. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*. 5:e8.
7. Meyer PE, Kontos K, Lafitte F, Bontempi G (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinf Syst Biol* 2007:79879.
8. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121.
9. Martínez-Antonio A, Janga SC, Salgado H, Collado-Vides J (2006) Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol* 14:22–7.
10. Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genetics* 5:34–42.
11. Ulrich LE, Koonin EV, Zhulin IB (2005). One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* 13:52–6.
12. Kashtan N, Alon U (2005) Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci*. 102:13773.
13. Kirkpatrick S, et al. (1983) Optimization by simulated annealing. *Science* 220:671–680.
14. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 106:620–630.
15. Ma W, et al. (2009) Defining network topologies that can achieve biochemical adaptation. *Cell* 138:760–773.
16. Conrad TM, et al. (2009) Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol* 9:R118 .
17. Isalan M, et al. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452:840–845.
18. Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. *Nat Rev Genet* 5:101–113.

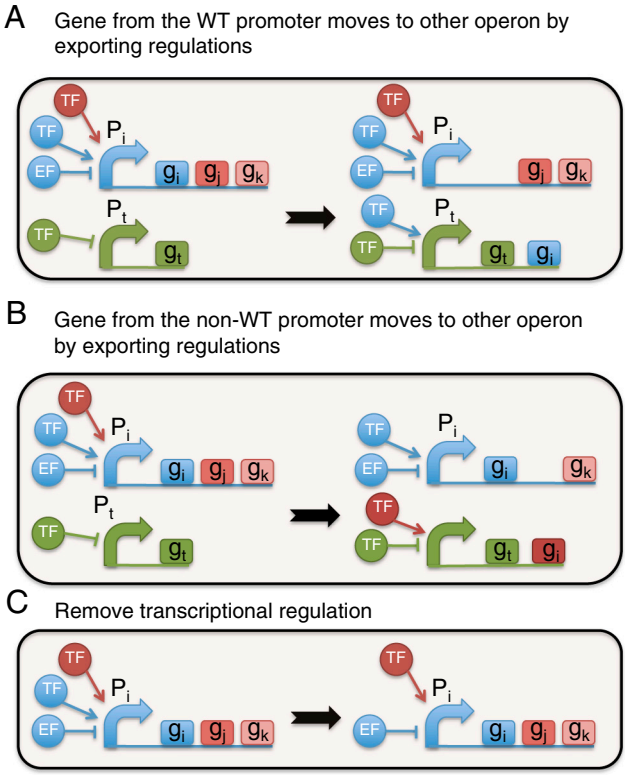

Fig. S1. Transcriptional mutations in the GTRN imposed during the evolutionary process (*Genome-Wide Optimization Procedure*, *SI Materials and Methods*).

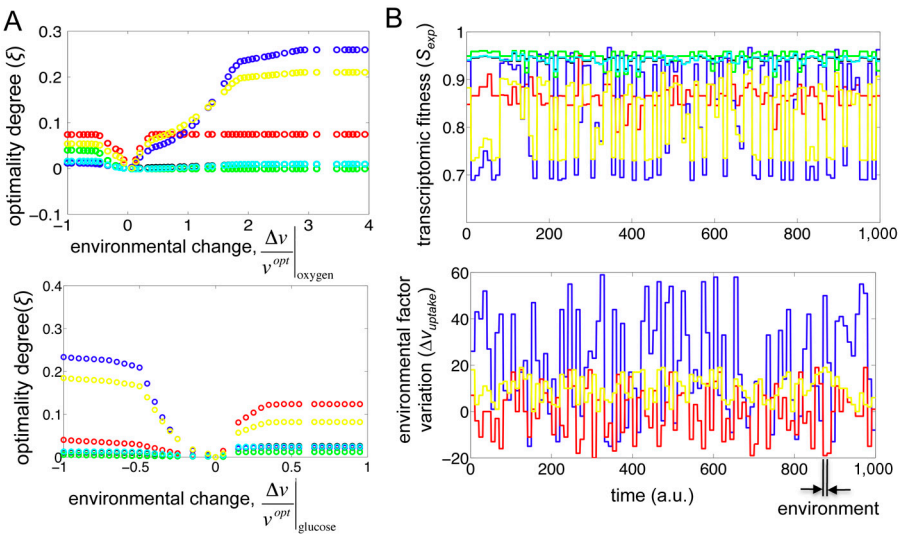

Fig. S2. (*A*) Optimality of the WT GTRN under environments perturbed in a single direction (oxygen and glucose). Notice that each simulation considered different sets of genes related to adaptation functions (red line), defense pathways (blue line), a combination of genes related to adaptation and defense (yellow line), protection (green line), central metabolism (cyan), and all genes (black line). (*B*) Quantitative simulations of transcriptomic fitness, S_{exp}, under a variety of environmental conditions, Δv_k, simultaneously (colors in *A* were maintained to indicate the different sets of genes selected to evaluate S_{exp} in *B*, *Top*). At the bottom of *B*, the external fluxes of glucose (red line), NO_3^- (yellow line) and oxygen (blue line) represent changes in carbon and nitrogen availability and the cellular uptake of oxygen, respectively. Each environment is defined using random values for external oxygen flux and carbon and nitrogen availability that range from minimal to saturating values, thereby simulating extreme environments.

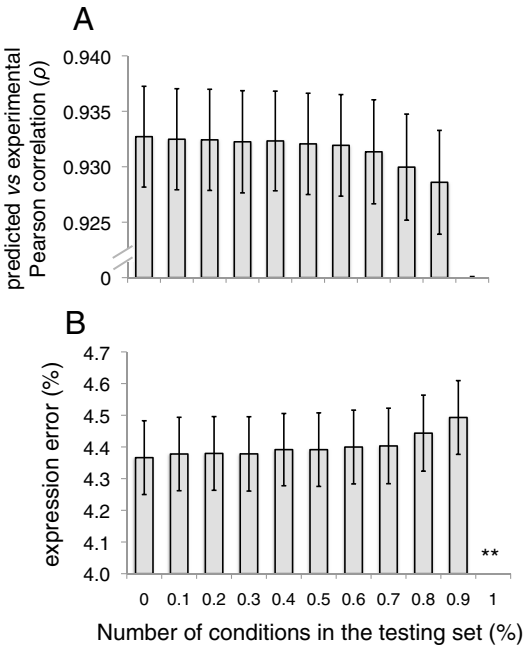

Fig. S3. Predicted versus experimental Pearson correlation, $\rho(A)$, and gene expression error, $e(B)$, computed in testing sets with different sizes that contain experimental conditions of M3D selected randomly. Note that each model was trained by the conditions included in M3D excluding the testing set. SD represents are related to the scores in the 380 conditions.

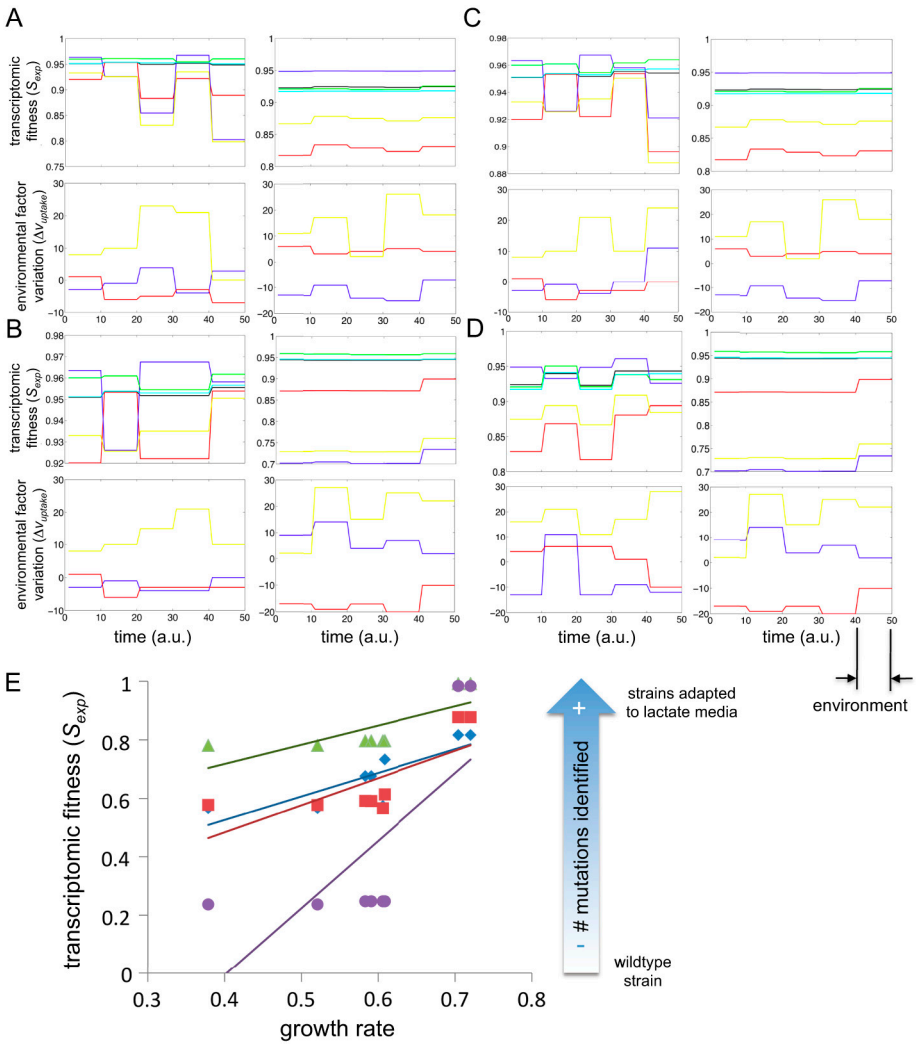

Fig. S4. Expression score of the *E. coli* WT GTRN under several sets of environments defined by different levels of environmental fluxes with minimum (left figures) and maximum (right figures) variations of , considering only genes coding for central metabolism (*A*), adaptation- and defense-related genes (*B*), all genes (*C*) and defense-related genes (*D*) when evaluating S_{exp}. For all panels, we plotted all S_{exp}, defined by selecting target genes with the corresponding biological functions. (*E*) Correlations between predicted fitness considering only TFs (blue), genes coding for enzymes (green), genes related to stress (violet), or all genes (red) and the growth rate of four strains and their intermediaries that were evolved in the laboratory in minimal lactate medium.

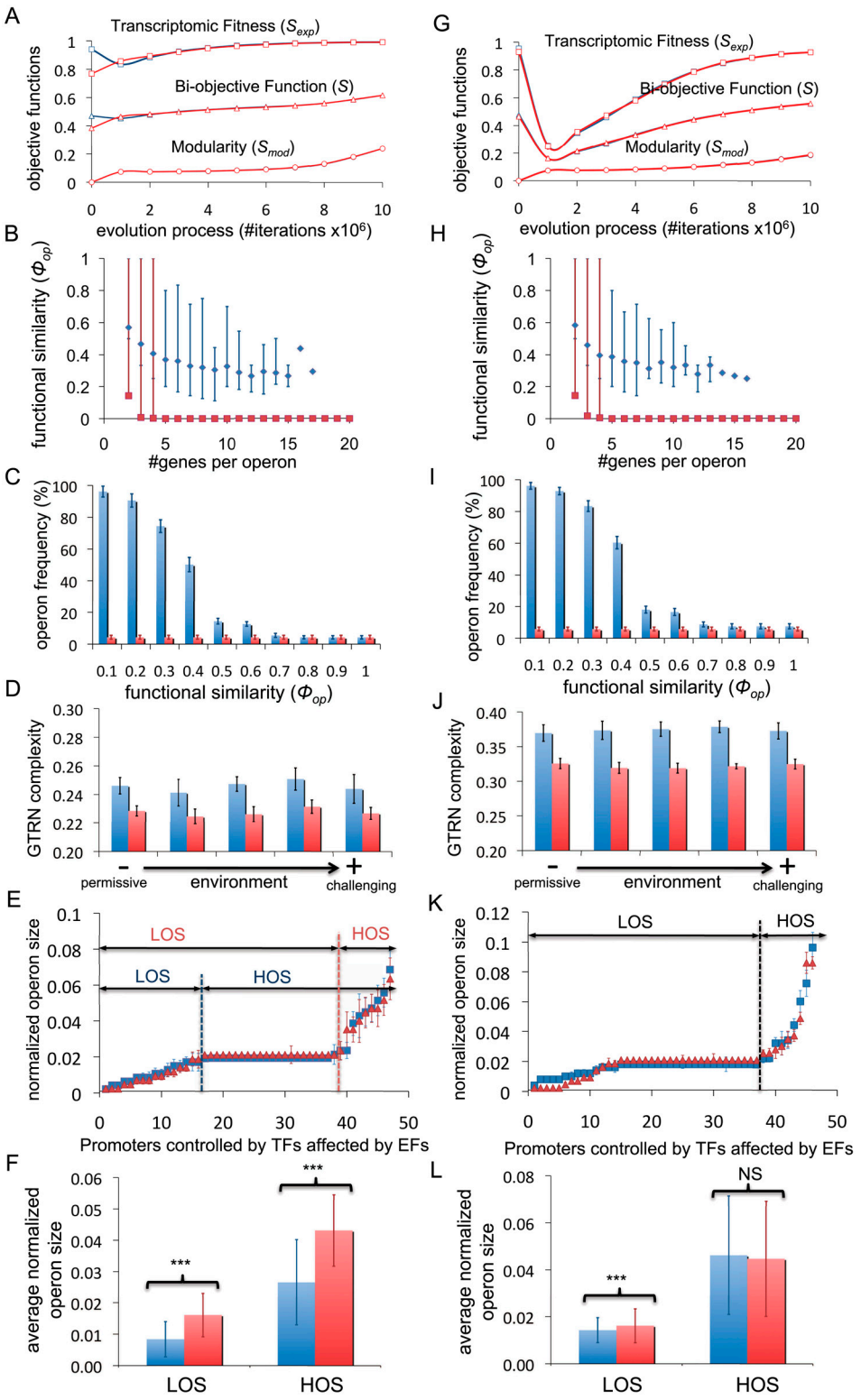

Fig. S5. Rewired GTRNs under selective pressure based on the expression of stress-related genes involved in adaptation and defense (*A–F*) or the entire genome (*G–L*). (*A* and *G*) Transcriptomic fitness and modularity score, and the bi-objective function for rewired GTRN in the evolutionary steps. Weighting factors were equal for both objectives. Random optimization produced significantly lower bi-objective function values than those observed for the WT GTRN. (*B* and *H*) Functional similarity, depending on operon size, for the rewired GTRNs that evolved in environment 3. (*C* and *I*) Histogram showing the functional similarity scores of the operons in rewired GTRNs under permissive environments. (*D* and *J*) Complexity reduction (the number of regulators and operons are represented using blue and red bars, respectively) of the rewired GTRNs with respect to WT GTRN under different levels of environmental extremity (these range from the most permissive to the most constrained) (*SI Materials and Methods*). Operons with size equal to one (null-similarity) were not plotted in *C*, *D*, *I*, and *J*). (*E* and *K*) Sizes of operons containing one or more genes controlled by promoters that interact with EFs in the rewired GTRNs under permissive and harsh environments (blue squares and red triangles, respectively). The vertical dashed line indicates the mean operon size value. Low operon size (LOS) and high operon size (HOS) classes were defined using the average operon size as a cutoff. (*F* and *L*) Mean operon size in rewired GTRNs under weak and strong environments (blue and red bars, respectively) for LOS and HOS classes. The *U*-test significance is shown (***$p < 0.01$; NS: not significant). Error bars represent the standard deviations obtained from 10 in silico evolution experiments.

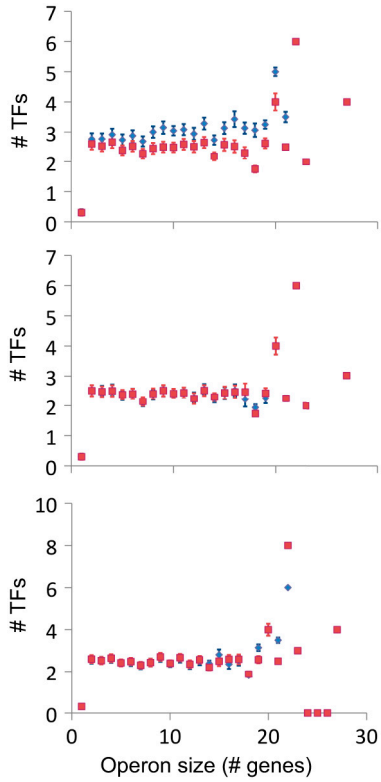

Fig. S6. Number of regulators (rewired and WT GTRNs) in promoters controlling operons of different sizes. Operons in rewired GTRNs (with selective pressure on the expression of enzymes, stress genes, or the whole genome, from top to bottom, respectively) under permissive and challenging environments are represented by blue and red points, respectively. Error bars represent mean standard errors of the number of regulators for each operon contained in rewired GTRNs obtained from 10 evolutionary processes.

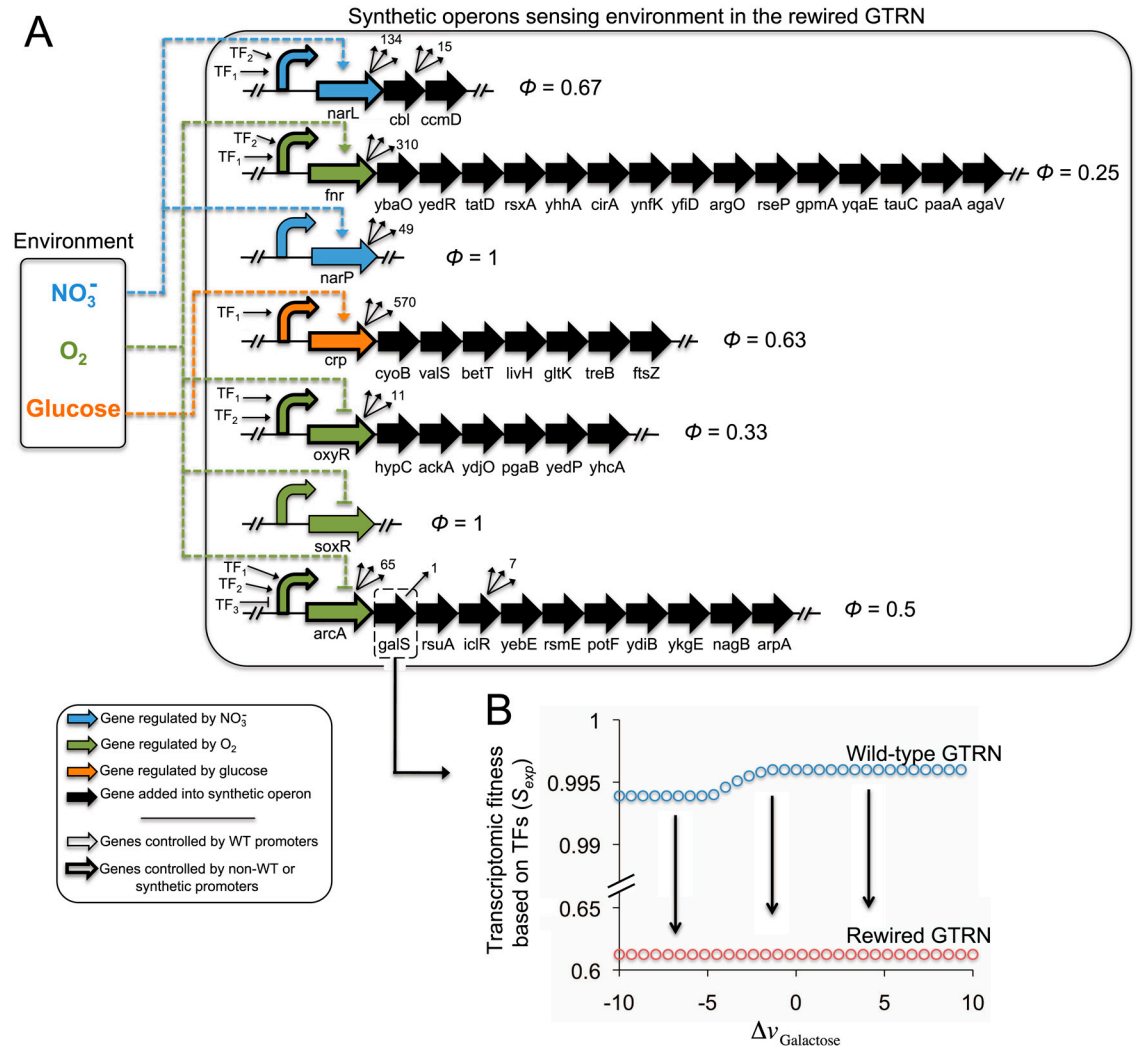

Fig. S7. (*A*) Synthetic operons of the rewired GTRN showed in Fig. 4*C*, *Lower*, sensing EFs. Note that we show the functional similarity (ϕ) of each operon. (*B*) Transcriptomic fitness in the WT and rewired GTRN under environments with different levels of galactose sensed by *galS*. Rewired GTRN was obtained imposing a selective pressure based on TFs and evolving in permissive environments.

Table S1. Topological parameters of the WT transcription network of *E. coli* and the rewired GTRN evolved in permissive and challenging environments. GTRN, genomic transcriptional regulatory network; TF, transcription factor; TRN, transcriptional regulatory networks

Parameter	WT	I*	II*	I*	II*
Clustering coefficient	0.103	0.021	0.028	0.019	0.021
Network diameter	9	15	16	14	14
Characteristic path length	2.897	4.952	5.098	4.724	4.500
$\gamma_{incoming}$ †	0.82	0.687	0.689	0.673	0.674
$\gamma_{outcoming}$ †	2.79	2.217	2.275	2.270	2.360
γ_{cc} ‡	0.98	0.843	0.757	0.782	0.831
Genes regulated by one TF	1122 (26%)	587 (14%)	592 (14%)	807 (19%)	801 (19%)
Genes regulated by two TFs	603 (14%)	770 (18%)	796 (19%)	687 (16%)	741 (18)
Genes regulated by three TFs	380 (9%)	513 (12%)	466 (11%)	472 (11%)	501 (12%)
Genes regulated by more than four TFs	509 (12%)	716 (17%)	756 (18%)	637 (15%)	641 (15%)
Synthetic promoters	—	237 (35%)	242 (17%)	126 (21%)	118 (20%)
Synthetic regulations	—	327 (9.5%)	326 (17%)	173 (11%)	164 (11%)

*Rewired GTRN with a selective pressure that takes into account gene expression of genes with enzymatic activity or related with adaptation or defense processes, respectively.

†Exponents of the power-law that fits the incoming ($r > 0.84$) and outgoing ($r > 0.85$) connectivity distributions of the TRN, respectively.

‡Slope ($r > 0.63$) of the linear regression between connectivity of each gene and its clustering coefficient in log-log scale.

Other Supporting Information Files

Dataset S1.

Signal transduction model parameters. (*Data S3*) Predicted transcriptional and environmental perturbations.

Dataset S1 (XLS)