Supporting Information

Burda et al. 10.1073/pnas.1109435108

SI Text

Detailed Specification of the Model. ***The transcriptional dynamics.*** The specification of the transcriptional dynamics requires including thermodynamic binding of transcription factors (TF) to their binding sites and modeling the consequences on the rate of transcription. Consider first the case where all N transcription factors affect gene i as activators. If at least one of the binding sites is occupied by its TF, we consider that the gene will be transcribed; this choice corresponds to having the transcription rate be proportional to the Probability of OCCupation or "POCC" (1) of the regulatory region. In the simplest framework, we take the bindings to arise independently (no cooperativity). In addition, we identify up to an overall scale the transcription rate of a gene with the POCC of its regulatory region. Considering that protein content is proportional to transcription rate (at least in the steady-state), we set S_i—the mean normalized expression level of gene i—equal to this POCC. Using Eq. **1** of the main part of the paper, one then obtains

$$S_i = 1 - \prod_j (1 - P_{ij}). \qquad \textbf{[S1]}$$

This equation is basic for the "mean field" model of ref. 2 in which the neglect of fluctuations and the corresponding limitations were explained. Note that if the P_{ij} are small, transcription is additive in these variables, while in the binary limit where P_{ij} is 0 or 1, S_i corresponds to the logic of transcription being "on" if and only if at least one of the binding sites is occupied, as expected from the use of the POCC.

Our treatment of *inhibitory* interactions (due to repressors) is new and is motivated by a number of known cases where the binding of a TF acts as a veto, for example if the presence of the TF makes the DNA form a loop that conceals the other binding sites. Another mechanism for vetoing transcription is simply for the bound TF to block access of the polymerase to its promoter. Within our framework, transcription proceeds as in Eq. **S1** in the absence of any repressors, but as soon as any of the inhibitory sites are bound by their repressors, transcription is turned off. Again assuming there are no cooperative effects, and repeating for repressors the argument just used for activators, we are led to modify Eq. **S1** to

$$S_i = [1 - \prod_j (1 - P_{ij})] \prod_{j'} (1 - P_{ij'}), \qquad \textbf{[S2]}$$

where j runs over activating interactions and j' over inhibitory interactions.

By neglecting cooperative effects, we have obtained a model where the main parameters are those determining the binding probabilities implicit in Eq. **1** in the main part of the paper and these are subject to experimental constraints. All of our results are given for this model. Incorporating cooperative effects could lead to a more realistic model but at the cost of more parameters. We now propose a way to generalize the framework described so far. Reconsider the POCC of gene i's regulatory region; we denote by P_i this probability. Assuming that

$$P_i = \sum_{k \geq 1} \sum_{[j_1,\ldots,j_k]} P_i^{(k,N-k)}(j_1,\ldots,j_k,\bar{j}_{k+1},\ldots,\bar{j}_N), \qquad \textbf{[S3]}$$

where j_l ($\bar{j}_l$) is the label of an occupied (unoccupied) binding site and $[j_1,\ldots,j_k]$ stands for a combination of k out of N gene labels. For pedagogical reasons, reconsider the case where the bindings arise independently (no cooperativity). Then the probabilities in the sum on the right hand side would factorize into a product of terms $P_i^{(1,0)}(j)$ (or $P_i^{(0,1)}(\bar{j})$). Replacing $P_i^{(1,0)}(j)$ by P_{ij} defined in Eq. **1** of the main part of the paper, we have

$$P_i^{(k)}(j_1,\ldots,j_k,\bar{j}_{k+1},\ldots,\bar{j}_N) = \prod_j P_{ij} \times \prod_{j'} (1 - P_{ij'}), \Lambda \qquad \textbf{[S4]}$$

where j runs over indices for which the binding site is occupied and j' runs over the other indices. Then the sum over k in Eq. **S3** can be explicitly performed. However, to obtain a generalization of these equations so as to allow nonindependent binding probabilities, one could for instance replace in Eq. **S4** the equality by a proportionality. Such an identification often appears in the literature: using the stationary limit of appropriate kinetic equations, one argues that the concentration of a molecular complex is proportional to the product of concentrations of the constituents. Here, because of the reparametrization symmetry of the dynamics, the proportionality constant can only depend on k. One could then truncate the sum over k, say at $k = 3$, to avoid too many free parameters, a situation that arises in a number of genetic network reverse engineering attempts. Such a model deserves study, but that would take us much beyond the scope of the present work.

Defining phenotypes. The genotype of the GRN is its *hardware*, specified by the list of interaction weights W_{ij} themselves determined by the mismatches between the character strings associated with TF and binding sites. By convention, we make the weight negative for a repressor and positive for an activator. The phenotype of a given GRN is associated with its expression behavior which follows from the transcriptional dynamics. We consider two cases of behavior. In the first, we focus on the steady-state expression vectors (fixed points of the transcriptional dynamics). In the second, we focus on cyclic behavior of the expression vectors.

Given a GRN genotype, determining its phenotype is straightforward in practice. In the first case where we have given target expression patterns, we start in these target vectors and we see whether we converge to a nearby fixed point under iteration of the transcriptional dynamics. (In contrast, in our previous work, we had considered initial states that were unrelated to the target vector.) In the second case, we start with one of the patterns in the target cycle and see whether the trajectory under iterations stays close to that cycle. For the steady-state behavior, we shall impose 2, 3, or more vectors that consist of $N/2$ levels at 0 and $N/2$ at 1, and furthermore these vectors are taken to be orthogonal (for the 0/1 coding for S_i this means that the scalar product of two vectors is $N/4$). Setting $N = 16$ (the choice of N is not important as long as it has a moderate value, we have not explored what happens at large N), we define four mutually orthogonal targets as follows, a direct generalization of that of ref. 2:

1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0

This choice is motivated by the fact that at large N, random binary vectors are typically nearly orthogonal. The symmetries of the

model are relevant for studying it, but with any interesting choice of vectors, most of the symmetries are broken. Note that because one can transform these vectors into each other by index permutation, the basins of attraction leading to these targets are on average of equal size.

For the case where one enforces a target cycle, we shall use the toy sequence where the genes are taken to lie on a ring, and the cycle consists in having the "on" genes shift to the right at each time step:

1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0
0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1
1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0

The above is reminiscent of the cycle studied by Li et al. (3) for the yeast cell cycle. We have also considered a cycle where the shift is not by two, but by one or three genes. The results are nearly the same in all these cases.

As a last point, when initializing the expression vector to determine the phenotype, instead of zero values we use a very small nonzero value (e.g., 10^{-3}) so that the target expression is an attractor which is in fact the desired property for the phenotype to be relevant.

A fitness for deviations from the target behavior. To quantify the deviations from the ideal target behavior, we first check whether we have steady-state behavior (in the first case) or cyclic behavior (in the second). For each target vector $\mathbf{S}^{(\text{target})}$, we define its distance to the associated GRN specific expression vector $\mathbf{S}$ via:

$$D(\mathbf{S},\mathbf{S}^{(\text{target})}) = \sum_i |S_i - S_i^{(\text{target})}|. \qquad \textbf{[S5]}$$

By summing all these distances, one for each target (each steady-state in the first case, and each expression vector of the periodic cycle in the second), we obtain what we refer to as the "total" distance D_T for that GRN. The resulting measure of "fidelity" to the imposed function can be turned into a kind of fitness via

$$F(\text{GRN}) = \exp(-fD_T), \qquad \textbf{[S6]}$$

where f acts as a control parameter allowing one to be more or less stringent on the fidelity. We thus consider the set of all GRN and apply the relative weight $F(\text{GRN})$ to each; this then provides an *ensemble* for the GRN, and by adjusting f we can focus on those GRN that are the most functional. For specificity, we shall work with $f = 20$, but our results depend only very weakly on this choice provided f is in the range 10 to 100.

Computational Approach via MCMC Sampling. A random genotype will not produce vectors close to the target ones because there are so many possibilities of expression vectors. Thus in practice one cannot usefully sample functional GRN by randomly generating character strings or equivalently random Ws. Instead, we note that the space of all genotypes is endowed with a fitness function (see previous paragraph) and thus one can apply Markov Chain Monte Carlo (MCMC) techniques. We use MCMC with the Metropolis rule (4). That computer algorithm produces a (biased) random walk in the fitness landscape that visits at long times the different genotypes according to their fitness as given in Eq. **S6**; the sampling thus focuses on genotypes having high fidelity to the imposed functional constraints. In detail, we perform random mutations of the binding sites, which produce changes to the edge weights and thus to the genotypes. (Technically, it would be possible to work at the level of edge weights alone, but it would make explanations of the MCMC far more delicate.) A *sweep* is defined as LN^2 successively attempted changes of the genotype (a random mutation of one coding letter and, independently, a random switch of the sign of one of the TF-DNA interactions). Each such change is accepted or rejected by the Metropolis algorithm. The Markov chain is initialized for instance using a random GRN. After some time, as in ref. 2, we obtain a GRN with a phenotype sufficiently close to the target (see Fig. S1), and we use this to start the production run of the MCMC. Thereafter, we iterate sweeps, recording successive GRNs.

A speed optimization was implemented to address the problem of CPU times: the simulation as described above requires a lot of computation time at $L = 12$ where small mismatches become very unprobable and one has to wait extremely long to go from one genotype to another one that is relatively independent. Therefore, we implemented the following modified procedure. We first set $L = 8$ and apply the MCMC, recording genotypes every 100 sweeps. Because our dynamics depend on mismatches and not on the value of L, these recorded GRNs can be interpreted as $L = 12$ genotypes with high fitness; the only issue is that the distribution of the magnitudes of the corresponding mismatches is incorrect. Hence we set $L = 12$ and use the saved genotypes obtained at $L = 8$ as starting points of new MCMC runs, obtaining from these after equilibration a large sample of $L = 12$ genotypes in a reasonable time. Notice that, in principle, any pairs of GRNs can be reached from one another via a succession of elementary moves. In this sense our sampling is ergodic.

Essential Interactions and the Essential Network. As already mentioned, genotypes can be represented by the N by N matrix of W_{ij} interaction strengths along with the N^2 signs specifying the activating vs. inhibitory nature of each interaction. These W_{ij} are never zero (see Eq. **2** of the main part of the paper), so we cannot say that an interaction is completely absent. Nevertheless, one may expect some interactions to be more important than others, for instance when the W_{ij}s are larger than average. An arbitrary cut-off could be introduced for separating small and large values, but it is better to base such a classification on functionality. We thus consider what happens when an interaction W_{ij} is removed by setting it to zero. Starting with one of the genotypes generated by our MCMC (and thus typically satisfying well the soft functional constraints), we determine the change in fitness produced by setting W_{ij} to zero: if the change is rejected by the Metropolis rule in five successive attempts, we say that this interaction is *essential*, motivated by the corresponding biological definition (a very similar result is obtained by defining the essentiality as the sensitivity to a single mutation that increases the mismatch by 1; because the definition of essentiality involves the Metropolis test, a random event, one sometimes finds false essentials, however this is a very weak effect). This definition leads to a *summary* description of a genotype via a list of ordered pairs (i,j) specifying the essential interactions as well as their nature (activating or inhibitory). This list of pairs can then be represented by a directed graph, with + signs on the edges that are activating and − signs on the edges that are inhibitory. In all our work we refer to this oriented and signed graph as the *essential network* of the genotype; note that no information on the weights W_{ij} is attached to this network representation.

SI Results. ***Basins of attraction for the transcriptional dynamics.*** We have checked that the basin of attraction of a target phenotype represents a large fraction of all possible initial phenotypes, so that the GRN property of performing the imposed function is

not merely a consequence of starting with the right expression vector. For instance, in the case of the fixed point phenotypes, these basins constitute approximately 99.8(9)%, 52.9(9)%, and 49.6(8)% of the whole space for $n = 2$, 3, and 4 respectively. With our choice of target phenotypes and the permutation symmetry of the model, the basins associated with individual targets are equal (after averaging over functional GRNs).

Statistics of essential interactions. The essential network provides a summary of the GRN genotype that can be more easily understood than the full genotype. A key feature of this summary is its sparseness: there are few essential interactions controlling any given gene. Another property is the relative frequency of inhibitory and activating interactions. For pedagogical reasons, let us first consider this issue at the level of all interactions, not just the essential ones. Even though the genotypes generated by the MCMC sampling have functional constraints, they contain many small W_{ij} that have hardly any effect on the phenotype; the sign of these interactions are thus random, and in effect these W_{ij} act as noise. If instead we focus on the larger W_{ij}, the functional constraints are likely to bias the sign in favor of activating interactions. To avoid an arbitrary definition of large weights, it is advantageous to use the notion of essentiality because of its link with phenotypes. For the essential networks produced from the GRN of our MCMC with the constraint of two to four steady states, we find that the great majority of the essential interactions are activating, see Table S1 (these numbers are not sensitive to N). These results are not surprising, but increasing the number of constraints forces the connections to be more complex and to make greater use of inhibitory interactions. In the toy cases of genes on a ring, we also have this general picture and find that the number of both activating and inhibitory essential interactions grows linearly with N.

Abundance of functional essential networks. Another question of interest concerns the number of distinct functional *essential networks*. The number of distinct GRNs is of little interest, being trivially enormous because all inessential interactions can be changed at will without affecting the phenotype. It is wise to first find the essential networks that are in a sense representative of a group of GRNs, in other words to perform a cluster analysis of the sample of essential networks at our disposal. Let the numbers of such networks be M and define a distance between a pair of them, for example

$$\text{Distance}(A,B) = \sum_{ij} (A_{ij} - B_{ij})^2, \qquad \textbf{[S7]}$$

where A_{ij}, B_{ij} are ± 1 for essential interactions and 0 otherwise. Our question can now be reformulated more precisely: does the number of clusters, considered as a proxy for the number of representative essential networks, saturate at some moderate value as M grows? (It must saturate somewhere, of course, but that may be for very large M values.) To answer this question, it is most convenient to use the modern affinity propagation algorithm (5), where the number of clusters is not preassigned but is determined by the algorithm; the code can be downloaded from Frey Laboratory Web page at Toronto University. As an illustration, for four fixed points we find that the number of clusters grows at large M roughly like $M^{2/3}$ (with a prefactor of the order of 0.3) and shows no sign of saturation up to at least $M = 4{,}000$. Other values of n lead to similar results, but some care is necessary in interpreting these trends at $n = 2$ and 3. Indeed, it turns out that for these values of n many clusters, distinct according to Eq. **S7**, have essentially the same topology and differ merely by the labeling of nodes (this reflects symmetries in our choice of the target phenotypes). In contrast, for $n = 4$ the clusters are genuinely different. To get more insight into this problem, we have carried out a complementary investigation, counting the number of distinct topologies (instead of using the clustering algorithm). This task is very tedious and our account of the network reparametrizations was only partial. With this proviso, it appears that the number of distinct topologies again increases like a power of M, however now the exponent increases with n (approximately from 0.69 for $n = 2$ to 0.97 for $n = 4$).

Motif frequency. The frequency of the most prominent motifs in our model is shown in Table S2.

Motifs using interaction strengths rather than essentiality. In our framework, every gene has some interaction with every other, so any analysis has to focus on the most relevant ones. So far, we have used the essentiality criterion, which keeps the functionally relevant interactions. What happens if instead we use the interaction strength as criterion? To address this question, we have recomputed the motif frequencies in the case of the cyclic phenotypes with a criterion based on the interaction's mismatch. Explicitly, we consider an interaction only if its mismatch is less or equal to $M_{\max}$. In Fig. S2 we show the dependence on $M_{\max}$. For small mismatches, the degree of overrepresentation of motifs is insensitive to $M_{\max}$, and so the result is nearly identical to that obtained when essentiality is used. However for larger allowed mismatches, we see that the overrepresentation of motifs drops sharply towards the background value frequency of randomized networks. This behavior is expected because allowing for large mismatches introduces W_{ij} that act like random noise.

A small oscillating GRN. To provide further insight into the properties of the transcriptional dynamics of our model, we consider here a small GRN with just three genes. We take the cyclic phenotype where each gene is turned on successively, with target expressions set to (1,0,0), (0,1,0), and (0,0,1). In Fig. S3 we show an example of a GRN produced and in which we show the successive expression levels when starting in the (1,0,0) configuration. In the absence of any input, a gene on its own will have a very low level of transcription within our model. To turn on during a cycling, a gene has to receive an activating signal from its predecessor. These signals are represented by the three red (activating) arrows in Fig. S3. [This feature differentiates our system from the repressilator network in which the genes are constitutively *on* (6).] In the presence of just these three activating interactions, the expression levels will end up all rising, approaching the (1,1,1) state. To prevent this saturation, multiple repressors are necessary. One way to achieve this goal is for all three genes to be self repressive and to repress their predecessor, but the MCMC finds more sparse solutions than that: here we see that two repressors are sufficient to have the desired cyclic behavior.

1. Granek J, Clark N (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology* 6:R87.
2. Burda Z, Krzywicki A, Martin OC, Zagorski M (2010) Distribution of essential interactions in model gene regulatory networks under mutation-selection balance. *Physical Review E* 82:011908.
3. Li F, Long T, Lu Y, Ouyang Q, Tang C (2004) The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci* 101:4781–4786.
4. Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equation of state calculations by fast computing machines. *J Chem Phys* 21:1087–1092.
5. Frey B, Dueck D (2007) Clustering by passing messages between data points. *Science* 315:972–976.
6. Elowitz M, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–338.

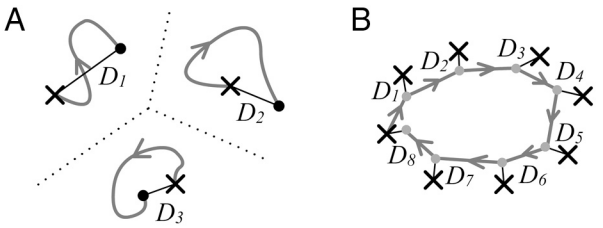

Fig. S1. A schematic representation of our MCMC process. (*A*) Steady-state behavior and $n = 3$: crosses (heavy dots) stand for the target (fixed point) states, while the line is the system's trajectory. The "total" distance entering the Metropolis test is $D_T = D_1 + D_2 + D_3$. (*B*) Similar as before, but for a cycle. Gray dots stand for successive states obtained by iterating Eq. **S2**. Here $D_T = D_1 + \ldots + D_8$.

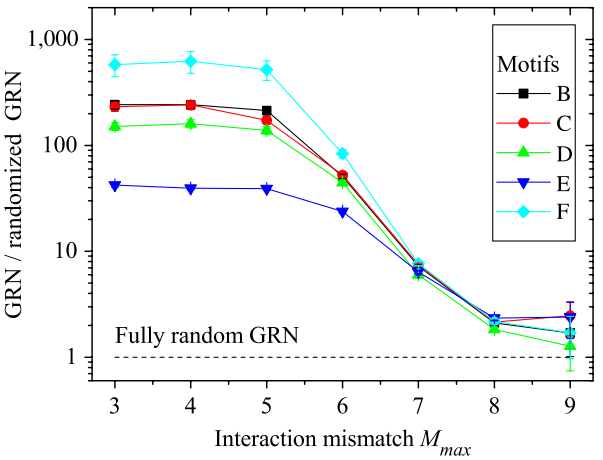

Fig. S2. Overrepresentation of motifs as a function of the interaction strengths allowed. The *x* axis gives the maximum mismatch $M_{\max}$ for an interaction to be included in the search for motifs. The *y* axis gives the frequency of each motif divided by the frequency in the randomized ensemble. We see that these ratios of frequencies are insensitive to $M_{\max}$ at low values but then rapidly decrease towards 1 at larger $M_{\max}$. ($N = 16$ and cyclic phenotypes are imposed.)

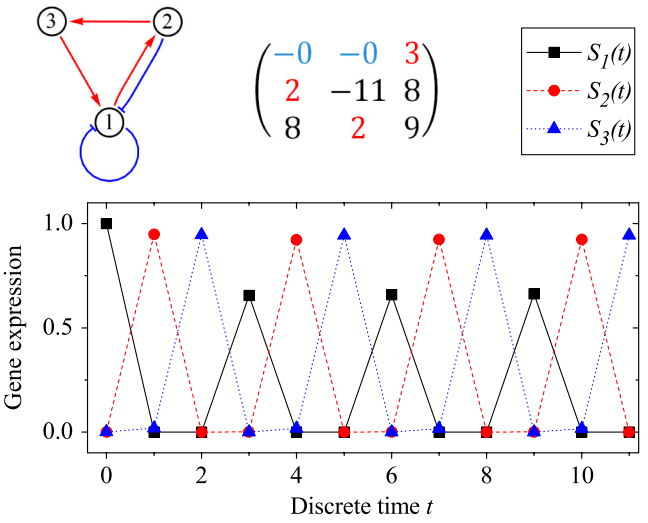

Fig. S3. A small system of three genes exhibiting oscillatory behavior. (Top) The network of essential interactions with the corresponding matrix of mismatches $\{d_{ij}\}$ (blue interactions and minus signs indicating repressors). (Bottom) The gene expression levels $S_i(t)$ produced by the transcriptional dynamics starting with the 1st gene being *on* and the other two being *off*.

Table S1. A sample of results of our simulations

Observable	2FP	3FP	4FP	Cycle
⟨#Essentials⟩	14.70(1)	21.08(2)	29.39(1)	23.42(1)
⟨#Inhibitory⟩	2.500(12)	6.38(2)	10.11(1)	7.38(1)
⟨Robustness⟩	0.9494(2)	0.9259(2)	0.8998(3)	0.9099(3)
⟨Dist2target⟩	0.966(3)	1.390(4)	1.961(4)	3.583(4)

nFP stands for phenotypes with n fixed points, while Cycle stands for the cyclic phenotype; for both, the number of genes is $N = 16$. The robustness is defined as the frequency of genotypes surviving a random mutation according to the Metropolis rule. Notice that it is fairly well reproduced by $1 - n/2N$, a result generalizing an analogous result of ref. 2. The distance to the target is the distance entering the fitness. It is nearly constant when divided by the number of target phenotypes. Further division by the number of active genes in a phenotype, i.e., $N/2$, yields approximately 6%, which measures the average deviation of their activity from the maximum $S_i = 1$.

Table S2. Most important motifs

Motif	2FP	3FP	4FP	Cycle
Motif *A*: model	**0.706(16)**	**2.358(39)**	**2.984(4)**	0.000
Randomized	0.002(1)	0.002(1)	0.002(2)	0.000
Motif *B*: model	0.000	0.000	0.000	**5.451(41)**
Randomized	0.001(1)	0.008(3)	0.071(9)	0.023(5)
Motif *C*: model	0.000	0.000	0.000	**5.170**(40)
Randomized	0.000	0.008(3)	0.078(9)	0.029(1)
Motif *D*: model	0.000	0.000	0.000	**4.533(42)**
Randomized	0.000	0.014(4)	0.0180(6)	0.030(7)
Motif *E*: model	0.000	0.000	0.000	**6.676(22)**
Randomized	0.017(4)	0.102(16)	0.057(8)	0.173(14)
Motif *F*:	0.000	0.000	0.000	**2.296(29)**
Randomized	0.000	0.001(1)	0.050(6)	0.003(2)

nFP means "n fixed points phenotype." Cycle refers to our 8-step cyclic phenotype. We typically used 1,000 GRNs in our motif search.