Adaptive evolution of feed-forward loops versus diamonds to filter out short spurious signals

Short title: Adaptive evolution of feed-forward loops and alternatives to them

Kun Xiong[1], Alex K. Lancaster[2], Mark L. Siegal[3], Joanna Masel[4*]

[1] Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona, United States of America

[2] Ronin Institute, Montclair, New Jersey, United States of America

[3] Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, United States of America

[4] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, United States of America

* Corresponding author

Email: masel@email.arizona.edu

Abstract

Transcriptional regulatory networks (TRNs) are enriched for certain network motifs. This could either be the result of natural selection for particular hypothesized functions of those motifs, or it could be a byproduct of mutation (e.g. of the prevalence of gene duplication) and of less specific forms of selection. We have developed a powerful new method for distinguishing between adaptive vs. non-adaptive causes, by simulating TRN evolution under different conditions. We simulate mutations to transcription factor binding sites in enough mechanistic detail to capture the high prevalence of weak-affinity binding sites, which can complicate the scoring of motifs. Our simulation of gene expression is also highly mechanistic, capturing stochasticity and delays in gene expression that distort external signals and intrinsically generate noise. We use the model to study a well-known motif, the type 1 coherent feed-forward loop (C1-FFL), which is hypothesized to filter out short spurious signals. We found that functional C1-FFLs evolve readily in TRNs under selection for this function, but not in a variety of negative controls. Interestingly, a new "diamond" motif also emerged as a short spurious signal filter. Like the C1-FFL, the diamond integrates information from a fast pathway and a slow pathway, but their speeds are based on gene expression dynamics rather than topology. When there is no external spurious signal to filter out, but only internally generated noise, only the diamond and not the C1-FFL evolves.

Author Summary

Frequently occurring motifs are thought to be fundamental building blocks of biological networks, conducting specific functions. However, we still lack definitive evidence that these motifs have evolved "adaptively" (to perform the particular function proposed for them), rather

than “non-adaptively” (as byproducts of some other function, or as an artifact of patterns of mutations). Here we develop a powerful null model that captures important non-adaptive factors that can shape the evolution of transcriptional regulatory networks, and use it to provide the missing piece of evidence of adaptive origin in the case of the most studied motif, a feed-forward loop that is hypothesized to filter out short spurious signals. We also find evidence for an alternative solution to this problem, where the functionality of the feed-forward loop is encoded not in network topology, but in the dynamics of gene expression. Our model is suitable for studying whether other network features have evolved adaptively vs. non-adaptively.

Introduction

Transcriptional regulatory networks (TRNs) are integral to development and physiology, and underlie all complex traits. An intriguing finding about TRNs is that certain "motifs" of interconnected transcription factors (TFs) are over-represented relative to random re-wirings that preserve the frequency distribution of connections [1, 2]. The significance of this finding remains open to debate.

The canonical example is the feed-forward loop (FFL), in which TF A regulates a target C both directly, and indirectly via TF B, and no regulatory connections exist in the opposite direction [1-3]. Each of the three regulatory interactions in a FFL can be either activating or repressing, so there are eight distinct kinds of FFLs [4; **Fig 1**]. Given the eight frequencies expected from the ratio of activators to repressors, two of these kinds of FFLs are significantly over-represented [4]. In this paper, we focus on one of these two over-represented types, namely the type 1 coherent FFL (C1-FFL), in which all three links are activating rather than repressing (**Fig 1**, top left). C1-FFL motifs are an active part of systems biology research today, e.g. they are used to infer the function of specific regulatory pathways [5, 6].

The over-representation of FFLs in observed TRNs is normally explained in terms of selection favoring a function of FFLs. Specifically, the most common adaptive hypothesis for the over-representation of C1-FFLs is that cells often benefit from ignoring short-lived signals and responding only to durable signals [3, 4, 7]. Evidence that C1-FFLs can perform this function comes from the behavior both of theoretical models [4] and of *in vivo* gene circuits [7]. A C1-FFL can achieve this function when its regulatory logic is that of an "AND" gate, i.e. both the direct path from A to C and the indirect path from A to B to C must be activated before the response is

76 triggered. In this case, the response will only be triggered if, by the time the signal trickles
77 through the longer path, it is still active on the shorter path as well. This yields a response to
78 long-lived signals but not short-lived signals.
79
80 However, just because a behavior is observed, we cannot conclude that the behavior is a
81 historical consequence of past selection favoring that behavior [8, 9]. The explanatory power of
82 this adaptive hypothesis of filtering out short-lived and spurious signals needs to be compared
83 to that of alternative, non-adaptive hypotheses [10]. The over-representation of C1-FFLs might
84 be a byproduct of some other behavior that was the true target of selection [11]. Alternatively,
85 it might be an intrinsic property of TRNs generated by mutational processes – gene duplication
86 patterns have been found to enrich for FFLs in general [12], although not yet C1-FFLs in
87 particular. Adaptationist claims about TRN organization have been accused of being just-so
88 stories, with adaptive hypotheses still in need of testing against an appropriate null model of
89 network evolution [13-23].
90
91 Here we develop such a computational null model of TRN evolution, and apply it to the case of
92 C1-FFL over-representation. We simulate gene duplication and deletion, and sufficient realism in
93 our model of cis-regulatory evolution to capture the non-adaptive effects of mutation in shaping
94 TRNs. In particular, we consider “weak” TF binding sites (TFBSs) that can easily appear de novo
95 by chance alone, and from there be selected to bind a TF more strongly.
96
97 It is also important to capture the stochasticity of gene expression, which causes the number of
98 mRNAs and hence proteins to fluctuate [24, 25]. This is because demand for spurious signal
99 filtering and hence C1-FFL function may arise not just from external signals, but also from

internal fluctuations. Stochasticity in gene expression also shapes how external spurious signals are propagated. Stochasticity is a constraint on what TRNs can achieve, but can be adaptively co-opted in evolution [26]; either way, it might underlie the evolution of certain motifs. Most computational models of TRN evolution that consider gene expression as the major phenotype do not simulate stochasticity in gene expression (see [27-29] for three notable exceptions). The genotype to phenotype map we develop here does include intrinsic stochasticity in gene expression.

Here we use this model to ask whether AND-gated C1-FFLs evolve as a response to selection for filtering out short and spurious external signals, compared to conditions that control for both mutational biases and for less specific forms of selection. We find that they evolve far more often under these specific selection conditions than under control conditions, providing long-awaited support for the adaptive hypothesis. We also ask whether there are alternative motifs that evolve to solve the same selective challenge. We find that a “diamond” [30] is such a motif, filtering out short spurious signals by requiring them to arrive not through both a long and a short path, but through both a fast and a slow path of equal topological lengths. We also compare motifs that evolve to filter out external spurious signals to those that evolve in response to intrinsic stochastic noise in gene expression. We find that while both diamonds and C1-FFLs evolve in response to the former, only diamonds evolve in response to the latter.

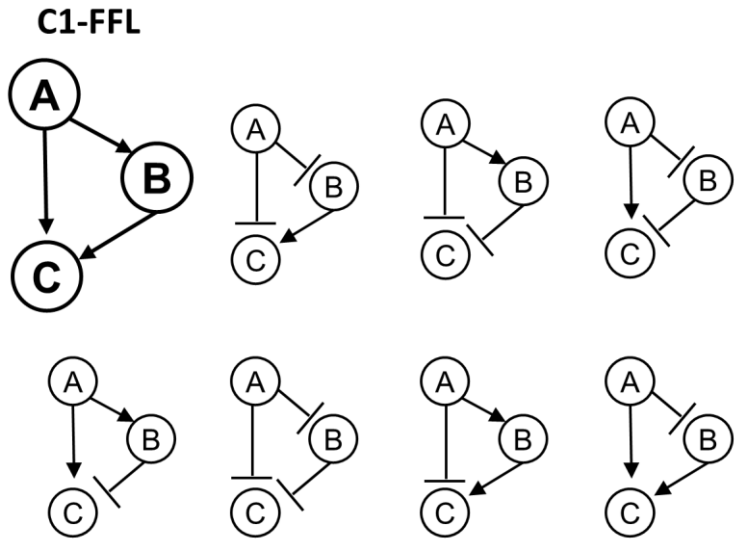


Fig 1. Feed-forward loops come in eight subtypes. TF A and TF B can activate (indicated by arrows) or repress (indicated by bars) expression of the effector C as well as other TFs. Auto-regulation is allowed, but not shown. Following Milo et al. [1], we exclude the case in which A and B regulate one another, rather than treating this case as two overlapping FFLs.

Models

Overview of the model

We simulate the dynamics of TRNs as the TFs activate and repress one another's transcription. For each moment in developmental time (i.e. on the timescale of one cell responding to stimuli), we simulate the numbers of nuclear and cytoplasmic mRNAs in a cell, the protein concentrations, and the chromatin state of each transcription start site. Transitions between three possible chromatin states -- Repressed, Intermediate, and Active -- are a stochastic function of TF binding, and transcription initiation from the Active state is also stochastic. An overview of the model is shown in **Fig 2**. The pattern of TF binding affects chromatin, which affects transcription rates, eventually affecting the concentration of TFs and so completing regulatory feedback loops. The genotype is specified by a set of cis-regulatory sequences that contain TFBSs to which TFs may bind (which, as nucleotide sequences, are subject to realistic mutational parameters), by which consensus sequence each TF recognizes and with what affinity, and by 5 gene-specific parameters that control gene expression as a function of TF binding: mean duration of transcriptional bursts, mRNA degradation, protein production, and protein degradation rates, and gene length which affects delays in transcription and translation. An external signal is treated like another TF, and the concentration of an effector gene in response is a primary determinant of fitness, combined with a cost associated with gene expression (**Fig 2**). Mutants replace resident genotypes as a function of the difference in estimated fitness. Parameter values, taken as far as possible from *Saccharomyces cerevisiae*, are summarized in **Table 1**. Source code in C is available at https://github.com/MaselLab/network-evolution-simulator.

148 **Table 1. Major model parameters**

Parameter	Values[1]	Bounds[2]	References
Length of cis-regulatory sequence	150 bp		[31]
Length of TF recognition sequence	8 bp		[32]
Length occupied by a TF on each side of recognition sequence	3 bp		[34]
Dissociation constant between TF and perfect TFBS, $K_d(0)$	**$10^{U(-9,-6)}$ M**[3]	$(0, 10^{-5})$	[37, 38]
Dissociation constant between TF and non-specific DNA, $K_d(3)$	10^{-5} M		[33]
Base rate of transition from Repressed to Intermediate	0.15 min^{-1}		[44]
Maximum transition rate from Repressed to Intermediate	0.92 min^{-1}		[40, 44]
Base rate of transition from Intermediate to Repressed	0.67 min^{-1}		[44]
Maximum transition rate from Intermediate to Repressed	4.11 min^{-1}		Chosen to give same dynamic range and Repressed to Intermediate
Base rate of transition from Intermediate to Active	0.025 min^{-1}		[40]
Maximum transition rate from Intermediate to Active	3.3 min^{-1}		[40]
Transition rate from Active to Intermediate, $r_{Act\_to\_Int}$	**$10^{N(1.27, 0.226)}$ min^{-1}**[4]	[0.59, 64.7]	[40, 49, 50]
Length of gene, L	**$10^{N(2.568, 0.34)}$ codons**	[50, 5000]	[79]
Rate of transcription initiation, $r_{max\_transc\_init}$	6.75 min^{-1}		[40]
Speed of transcription elongation	600 codon/min		[51, 80, 81]
Time for transcribing UTRs and for terminating transcription	1 min		[51, 80, 81]
Rate of mRNA degradation, r_{mRNA_deg}	**$10^{N(-1.49, 0.267)}$ min^{-1}**	$[7.5\times10^{-4}, 0.54]$	[82]
Speed of translation elongation	330 codon/min		[55]
Translation initiation time	0.5 min		[55]
Protein synthesis rate, $r_{protein\_syn}$	**$10^{N(0.322, 0.416)}$ molecule mRNA^{-1} min^{-1}**	$[4.5\times10^{-3}, 61.4]$	[55]
Rate of protein degradation, $r_{protein\_deg}$	**$10^{N(-1.88, 0.561)}$ min^{-1}**	$[3.0\times10^{-6}, 0.69]$	[83]
Saturation concentration of effector protein, N_{e_sat}	10,000 molecules/cell		[58]
Fitness cost of protein expression for a gene with $L = 10^{2.568}$, c_{transl}	2×10^{-6} $(molecules/min)^{-1}$		[58, 59]
Maximum number of effector gene copies	5		
Maximum number of TF gene copies, excluding the signal	19		

[1] Parameters in bold can be altered by mutation, and the table shows the distributions from which their initial values are sampled. Parameter estimation is described either in the Methods or **S1 Text** section 2-5.

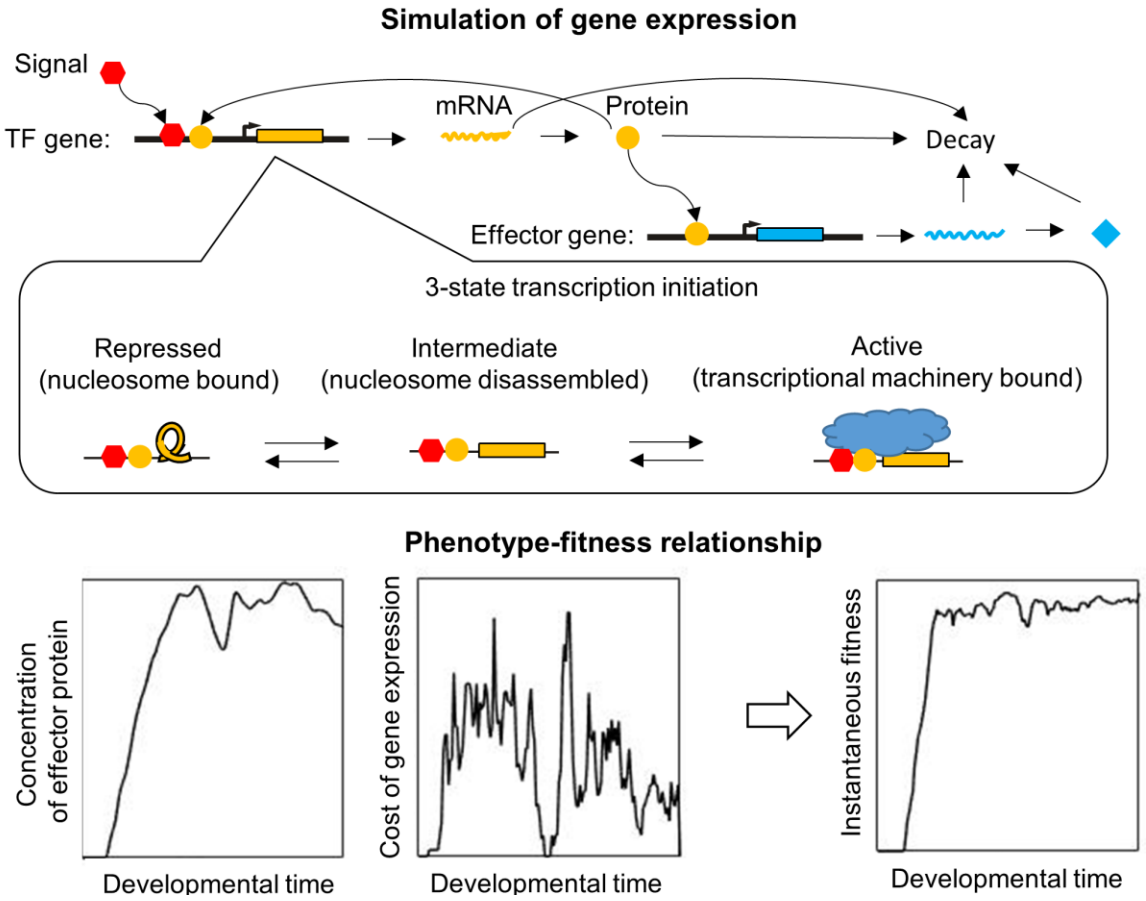
[2] Boundary values use the same unit as the parameter values. Parentheses mean the parameter cannot take the boundary values; brackets mean the opposite. We also use these bounds to constrain mutation (see **S1 Text** section 8).

[3] The uniform distribution is denoted as U(min, max).

[4] The normal distribution is denoted as N(mean, SD).

149

150



151

152 **Fig 2. Overview of the model.** As an example, we show a simple TRN that contains two genes.

153 Top: major biological processes (arrows) simulated in the model. Bottom: fitness is primarily

154 determined by the concentration of an effector protein (here shown as beneficial as in Eq. 2, but

155 potentially deleterious in a different environment as in Eq. 3), with a secondary component

156 coming from the cost of gene expression (proportional to the rate of protein production),

157 combined to give an instantaneous fitness at each moment in developmental time.

158

159 **Transcription factor binding**

160 Transcription of each gene is controlled by TFBSs present within a 150-bp cis-regulatory region,

161 corresponding to a typical yeast nucleosome-free region within a promoter [31]. The perfect

162 TFBS for a typical yeast TF has information content equivalent to 13.8 bits [32]; this means that

in a simplified model of binding where only one of the four nucleotides is a good match at each site, ~7 bp are recognized as an optimal consensus binding site. Maerkl & Quake [33] reported that the TFBSs of two yeast TFs, Pho4p and Cbf1p, can have up to 2 mismatched sites within their 6 bp consensus binding sequence, while still binding the TF above background levels [33]. Our model therefore tracks TFBSs with up to 2 mismatches. This low information content implies a higher density of TFBSs within our cis-regulatory regions than our algorithm was able to handle, so we instead assigned each TF an 8-bp consensus sequence. Two TFs cannot simultaneously occupy overlapping stretches, which we assume extend beyond the recognition sequence to occupy a total of 14 bp [34]; this captures competitive binding. Hindrance between TFBSs is shown in **Fig 3A**; TFs are assumed to work in both orientations [35].

Sites with m>3 mismatches are assumed to still bind at a background rate equal to m=3 mismatches, with dissociation constant $K_d(3) = 10^{-5}$ M [33] for all TFs. We assume that each of the last three bp makes an equal and independent additive contribution $\Delta G_{bp} < 0$ to the binding energy [36]: although not always true, this approximates average behavior well [33]. We ignore cooperativity in binding. Dissociation constants of eukaryotic TFs for perfect TFBSs can range from 10^{-5} M [37] to 10^{-11} M [38]. We initialize each TF with its own value of $\log_{10}(K_d(0))$ sampled from a uniform distribution between -6 and -9, with mutation capable of further expanding this range, subject to $K_d(0) < 10^{-5}$ M. Substituting m=0 and m=3 into

$$\Delta G_m = -RTlnK_d(m) = \Delta G_0 - \min(m, 3)\Delta G_{bp},$$

we can solve for ΔG_{bp} and ΔG_0, and thus obtain $K_d(1)$ and $K_d(2)$.

187 Because TFs bind non-specifically to DNA at a high background rate, each nucleosome-free
188 stretch of 14 bp can be considered to be a non-specific binding site (NSBS). A haploid *S.*
189 *cerevisiae* genome is 12 Mb, 80% of which is wrapped in nucleosomes [39], yielding
190 approximately 10^6 potential non-specific binding sites (NSBSs). In a yeast nucleus of volume
191 3×10^{-15} liters, the NSBS concentration is of order 10^{-4} M. To find the concentration of free TF
192 [TF] in the nucleus given a total TF concentration of C_{TF}, we consider
193
194 $$K_d = \frac{[\text{binding\_site}][\text{TF}]}{[\text{binding\_site} \cdot \text{TF}]},$$
195
196 in the context of NSBSs, substitute [TF·NSBS] with C_{TF} - [TF], and solve for
197
198 $$[\text{TF}] = \frac{K_d(3)}{K_d(3)+[\text{NSBS}]} C_{TF} = \frac{10^{-5}}{10^{-5}+10^{-4}} C_{TF}.$$
199
200 Thus, about 90% of total TFs are bound non-specifically, leaving about 10% free. The relatively
201 small number of specific TFBSs is not enough to significantly perturb the proportion of free TFs,
202 and so for the specific TFBSs with m<3 that are of interest in our model, we simply use $K_d^*(m)$ =
203 $10K_d(m)$ to account for the reduction in the amount of available TF due to non-specific binding.
204 We also rescale K_d^* from moles/liter to the more convenient number of molecules per cell by
205 multiplying by 3×10^{-15} liter $\times$ 6.02×10^{23} molecules/mole = 1.8×10^9 molecules cell^{-1} M^{-1}, for a
206 total multiplication factor of 1.8×10^{10} molecule M^{-1}. If there were only one binding site, it would
207 be bound for a fraction of time
208
209 $$P = \frac{N_i}{K_d^*+N_i} \qquad (1)$$

where N_i is the per-cell number of molecules of TF *i*; note that we assume all TF molecules are located in the nucleus.

The transition rates between chromatin states (see section below) are a function of the numbers of activators *A* and repressors *R* bound to a cis-regulatory region. Note that in our model, each TF is either always an activator, or always a repressor, independently of binding context. The joint probability distribution of *A* and *R* is derived in **S1 Text** section 1.

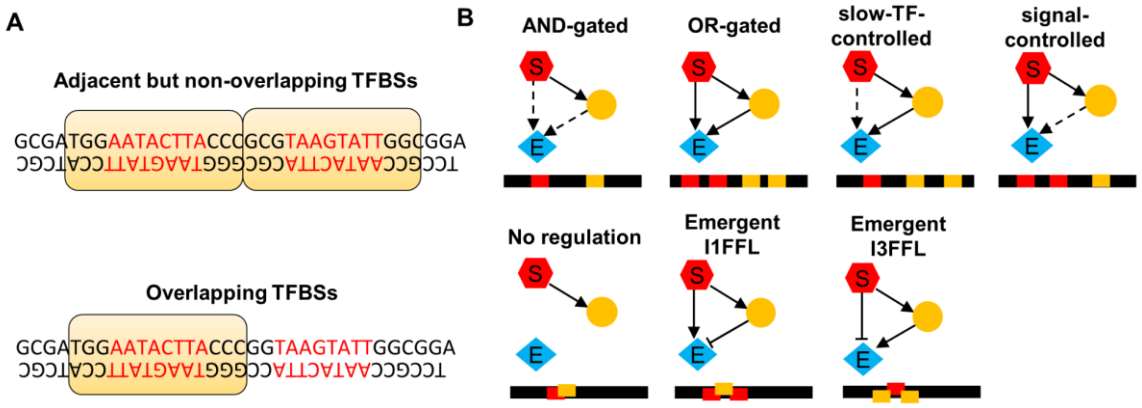


Fig 3. The numbers of TFBSs, and any hindrance between them, determines the regulatory logic of effector expression. (A) TFs (yellow boxes) recognize 8 bp (red) sites while occupying and thus excluding other TFs from a 14 bp long space. The sequence on the top allows simultaneous binding but that on the bottom does not. **(B)** We use the pattern of TFBSs (red and yellow bars along black cis-regulatory sequences) to classify the regulatory logic of the effector gene. C1-FFLs are classified first by whether or not they are capable of simultaneously binding the signal and the TF (top vs bottom). Further classification is based on whether either the signal or the TF has multiple non-overlapping TFBSs, allowing it to activate the effector without help from the other (solid arrow). The three subtypes on the bottom (where the signal and TF cannot

bind simultaneously) are rarely seen, and omitted from further analysis; they are shown here for completeness. I1-FFL and I3-FFL stand for type 1 and type 3 incoherent feed-forward loops, respectively [7].

Transcriptional regulation

Activation of the effector gene requires at least two TFBSs to be occupied by activators – not necessarily different activators. The requirement for two activators makes the effector gene capable of evolving an AND-gate via a configuration of TFBSs in which the only way to have two TFs bound is for them to be different TFs (**Fig 3B**). All other genes are AND-gate-incapable, meaning that their activation requires only one TFBS to be occupied by an activator. P_A denotes the probability of having at least one activator bound for an AND-gate-incapable gene, or two for an AND-gate-capable gene. P_R denotes the probability of having at least one repressor bound.

Noise in yeast gene expression is well described by a two step process of transcriptional activation [40, 41], e.g. nucleosome disassembly followed by transcription machinery assembly. We denote the three possible states of the transcription start site as Repressed, Intermediate, and Active (**Fig 2**). Transitions between the states depend on the numbers of activator and repressor TFs bound (e.g. via recruitment of histone-modifying enzymes [42, 43]). We make conversion from Repressed to Intermediate range, as a function of P_A, from the background rate 0.15 min^{-1} of histone acetylation [44; presumed to be followed by nucleosome disassembly], to the rate of nucleosome disassembly 0.92 min^{-1} for the constitutively active PHO5 promoter [40]:

$$r_{Rep\_to\_Int} = 0.92P_A + 0.15(1 - P_A).$$

We make conversion from Intermediate to Repressed a function of P_R, ranging from a background histone de-acetylation rate of 0.67 min^{-1} [44], up to 4.11 min^{-1}, with that maximum chosen so as to keep a similar maximum:basal rate ratio as that of $r_{Rep\_to\_Int}$:

$$r_{Int\_to\_Rep} = 4.11P_R + 0.67(1 - P_R).$$

We assume that repressors disrupt the assembly of transcription machinery [45] to such a degree that conversion from Intermediate to Active does not occur if even a single repressor is bound. In the absence of repressors, activators facilitate the assembly of transcription machinery [46]. Brown et al. [40] reported that the rate of transcription machinery assembly is 3.3 min^{-1} for a constitutively active PHO5 promoter, and 0.025 min^{-1} when the Pho4 activator of the PHO5 promoter is knocked out. We use this range to set

$$r_{Int\_to\_Act} = 3.3P_{A\_no\_R} + 0.025P_{notA\_no\_R}$$

where $P_{A\_no\_R}$ is the probability of having no repressors and either one (for an AND-gate-incapable gene) or two (for an AND-gate-capable gene) activators bound, and $P_{notA\_no\_R}$ is the probability of having no TFs bound (for AND-gate-incapable genes) or having no repressors and not more than one activator bound (for AND-gate-capable genes).

The promoter sequence not only determines which specific TFBSs are present, but also influences non-specific components of the transcriptional machinery [47, 48]. We capture this via gene-specific but TF-binding-independent rates $r_{Act\_to\_Int}$ with which the machinery

277 disassembles and a burst of transcription ends. In other words, we let TF binding regulate the
278 frequency of "bursts" of transcription, while other properties of the cis-regulatory region
279 regulate their duration. E.g., yeast transcription factor Pho4 regulates the frequency but not
280 duration of bursts of PHO5 expression, by regulating the rates of nucleosome removal and of
281 transition to but not from a transcriptionally active state [40]. We estimate the distribution of
282 $r_{Act\_to\_Int}$ from the observed rates of mRNA production of 255 yeast genes [49] that are likely to
283 have similarly low nucleosome occupancy [50] and thus are constitutively open to expression
284 (see **S1 Text s**ection 2 for details and also for the bounds of $r_{Act\_to\_Int}$). For modeling simplicity, we
285 assume that the core promoter sequence responsible for the value of $r_{Act\_to\_Int}$ is distinct from
286 the 150-bp sequences in which our TFBSs are found.
287
288 **mRNA and protein dynamics**
289 Once in the Active state, a gene initiates new transcripts stochastically at rate $r_{max\_transc\_init}$ = 6.75
290 mRNA/min [40]. There is a delay before transcription is completed, of duration 1 + L / 600
291 minutes, where L is the length of the ORF in codons (see **S1 Text** section 3).
292
293 We model a second delay between the completion of a transcript and the production of the first
294 protein from it. The delay comes from a combination of translation initiation and elongation; it
295 ends when the mRNA is fully loaded with ribosomes all the way through to the stop codon and
296 the first protein is produced. We ignore the time required for mRNA splicing; introns are rare in
297 yeast [51]. mRNA transportation from nucleus to cytosol, which is likely diffusion-limited [52,
298 53], is fast even in mammalian cells [54] let alone much smaller yeast cells, and the time it takes
299 is also ignored. The median time in yeast for initiating translation is 0.5 minute [Table 1 in 55],
300 and the genomic average peptide elongation rate is 330 codon/min [55]. After an mRNA is

produced, we therefore wait for 0.5 + L / 330 minutes, and then model protein production as continuous at a gene-specific rate $r_{protein\_syn}$ (see **S1 Text** section 4 for details of $r_{protein\_syn}$).

Protein transport into the nucleus is rapid [56] and is approximated as instantaneous and complete, so that the newly produced protein molecules immediately increase the probability of TF binding. Each gene has its own mRNA and protein decay rates, initialized from distributions taken from data (see **S1 Text** section 5).

All the rates regarding transcription and translation are listed in **Table 1**, including distributions estimated from data, and hard bounds imposed to prevent unrealistic values arising during evolution.

Developmental simulation

Our algorithm is part-stochastic, part-deterministic. We use a Gillespie algorithm [57] to simulate stochastic transitions between Repressed, Intermediate, and Active chromatin states, and to simulate transcription initiation and mRNA decay events. Fixed (i.e. deterministic) delay times are simulated between transcription initiation and completion, and between transcript completion and the production of the first protein. Protein production and degradation are described deterministically with ODEs, and updated frequently in order to recalculate TF concentrations and hence chromatin transition rates. We initialize developmental simulations with no mRNA or protein (except for the signal), and all genes in the Repressed state. Details of our simulation algorithm are given in the **S1 Text** section 6.

Selection conditions

Filtering out short spurious signals is a special case of signal recognition more generally. In environment 1, expressing the effector is beneficial, and in environment 2 it is deleterious. We select for TRNs that take information from the signal and correctly decide whether to express the effector. In our control condition, the signal is "on" at a constant level when the effector is beneficial in environment 1, and off in environment 2. Fitness is a weighted average across these two environments. In our test condition (**Fig 4**), the signal is constantly on in environment 1 and briefly on (for the first 10 minutes) in environment 2 – selection is to ignore this short spurious signal. The signal is treated as though it were an activating TF whose concentration is controlled externally, with an "off" concentration of zero and an "on" concentration of 1,000 molecules per cell, which is the typical per-cell number of a yeast TF [58].

We make fitness quantitative in terms of a "benefit" $B(t)$ as a function of the amount of effector protein $N_e(t)$ at developmental time t. Our motivation is the scenario in which the effector protein directs resources from metabolic program I to II. When program II produces benefits,

$$B(t) = \begin{cases} b_{max} \frac{N_e(t)}{N_{e\_sat}}, & N_e(t) < N_{e\_sat} \\ b_{max}, & N_e(t) \geq N_{e\_sat} \end{cases}, \quad (2)$$

where b_{max} is the maximum benefit if all resources were redirected to program II, and N_{e_sat} is the minimum of amount of effector protein to achieve this. Similarly, when program I is beneficial,

347 $$B(t) = \begin{cases} b_{max} - b_{max}\frac{N_e(t)}{N_{e_{sat}}}, & N_e(t) < N_{e_{sat}} \\ 0, & N_e(t) \geq N_{e_{sat}} \end{cases}. \quad (3)$$

348

349 We set N_{e_sat} to 10,000 molecules, which is about the average molecule number of a

350 metabolism-associated protein per cell in yeast [58]. Without loss of generality given that fitness

351 is relative, we set b_{max} to 1.

352

353 A second contribution to fitness comes from the cost of gene expression $C(t)$ (**Fig 2, bottom**

354 **center**). We make this cost proportional to the total protein production rate. We estimate a

355 fitness cost of gene expression of 2×10^{-6} per protein molecule translated per minute, based on

356 the cost of expressing a non-toxic protein in yeast [59; see **S1 Text** section 7 for details].

357

358 We simulate gene expression for 90 minutes of developmental time (**Fig 4**), and calculate

359 "cellular fitness" in a given environment as the average instantaneous fitness ($B(t)$-$C(t)$) over

360 these 90 minutes. We consider environment 2 to be twice as common as environment 1 (a

361 "signal" should be for an uncommon event rather than the default), and take the appropriate

362 weighted average.

363

364

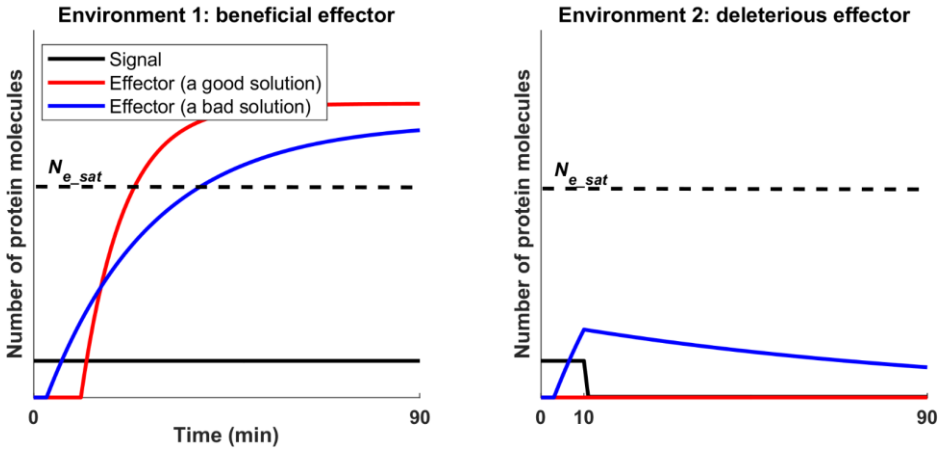


Fig 4. Selection for filtering out short spurious signals. The selection condition contains two environments. Each environment is a 90 min simulation of gene expression given signal input and the fitness effect of the effector. The signal is shown in black. Red illustrates favorable behavior of the effector in each of the environments, and, in comparison, blue shows a poor solution. See **S1 Fig** for examples of the evolved phenotypes.

Evolutionary simulation

We simulate a novel version of origin-fixation (weak-mutation-strong-selection) evolutionary dynamics, i.e. the population contains only one resident genotype at any time, and mutant genotypes are either rejected or chosen to be the next resident. Estimators $\hat{F}$ of genotype fitness are averaged over 200 developmental replicates per environment in the case of the mutant, plus an additional 800 should it be chosen to be the next resident. The mutant replaces the resident if

$$\frac{\hat{F}_{mutant} - \hat{F}_{resident}}{|\hat{F}_{resident}|} \geq 10^{-8}.$$

382 This differs from Kimura's [60] equation for fixation probability, but captures the same flavor;
383 due to stochasticity in $\hat{F}$, fixation probability is a monotonic function of the true difference in
384 fitness. Note that it is possible, especially at the beginning of an evolutionary simulation, for
385 relative fitness to be paradoxically negative. In this rare case, for simplicity, we use the absolute
386 value of $\hat{F}$ on the denominator.
387
388 If 2000 successive mutants are all rejected, the simulation is terminated; upon inspection, we
389 found that these resident genotypes had evolved to not express the effector in either
390 environment. We refer to each change in resident genotype as an evolutionary step. We stop
391 the simulation after 50,000 evolutionary steps; at this time, most replicate simulations seem to
392 have reached a fitness plateau (**S2 Fig**); we use all replicates except those terminated early. To
393 reduce the frequency of early termination in the case where the signal was not allowed to
394 directly regulate the effector, we used a burn-in phase selecting on a more accessible
395 intermediate phenotype (see **S1 Text** section 9). In this case, burn-in occurred for 1000
396 evolutionary steps, followed by the usual 50,000 evolutionary steps with selection for the
397 phenotype of interest (**S2 Fig**).
398
399 **Genotype Initialization**
400 We initialize genotypes with 3 activator genes, 3 repressor genes, and 1 effector gene. Cis-
401 regulatory sequences and consensus binding sequences contain As, Cs, Gs, and Ts sampled with
402 equal probability. Rate constants associated with the expression of each gene, are sampled from
403 the distributions described above and summarized in **Table 1**.
404
405 **Mutation**

A genotype is subjected to 5 broad classes of mutation, at rates summarized in **Table 2** and justified in **S1 Text** section 8. First are single nucleotide substitutions in the cis-regulatory sequence; the resident nucleotide mutates into one of the other three types of nucleotides with equal probability. Second are single nucleotide changes to the consensus binding sequence of a TF, with the resident nucleotide mutated into one of the other three types at equal probability. Both of these can affect the number and strength of TFBSs.

Table 2. Mutation rates and effect sizes

Mutation	Relative rate	Effect of mutation[1]
Single nucleotide substitution	5.25×10^{-8} per gene	
Gene deletion	1.5×10^{-7} per gene[2]	
Gene duplication	1.5×10^{-7} per gene[2]	
Mutation to consensus sequence of a TF	3.5×10^{-9} per gene	
Mutation to TF identity (activator vs. repressor)	3.5×10^{-9} per gene	
Mutation to $K_d(0)$	3.5×10^{-9} per gene	k = 0.5, μ = -5[2], σ = 0.776
Mutation to L	1.2×10^{-11} per codon	
Mutation to $r_{protein\_syn}$	9.5×10^{-12} per codon	k = 0.5, μ = 0.021[2], σ = 0.760
Mutation to $r_{protein\_deg}$	9.5×10^{-12} per codon	k = 0.5, μ = -1.88, σ = 0.739
Mutation to $r_{Act\_to\_Int}$	9.5×10^{-12} per codon	k = 0.5, μ = 1.57[2], σ = 0.773
Mutation to r_{mRNA_deg}	9.5×10^{-12} per codon	k = 0.5, μ = -1.19, σ = 0.396

Mutation to these quantitative rates takes the form $\log_{10}x' = \log_{10}x + \mathrm{Normal}(k(\mu - \log_{10}x), \sigma)$, where x is the original value of a rate and x' is the value after mutation. See **S1 Text** section 8 for details.

[2] The value of this parameter is different during burn-in. See **S1 Text** section 8 for details.

Third are gene duplications or deletions. Because computational cost scales steeply (and non-linearly) with network size, we do not allow effector genes to duplicate once there are 5 copies, nor TF genes to duplicate once the total number of TF gene copies is 19. We also do not allow the signal, the last effector gene, nor the last TF gene to be deleted.

Fourth are mutations to gene-specific expression parameters. Most of these (L, $r_{Act\_to\_Int}$, $r_{protein\_syn}$, r_{mRNA_deg}, and $r_{protein\_deg}$) apply to both TFs and effector genes, while mutations to the gene-specific values of $K_d(0)$ apply only to TFs. Each mutation to L increases or decreases it by 1 codon, with equal probability unless L is at the upper or lower bound. Effect sizes of mutations to the other five parameters are modeled in such a way that mutation would maintain log-

428 normal stationary distributions for these values, in the absence of selection or arbitrary bounds
429 (see **S1 Text** section 8 for details). Upper and lower bounds (**S1 Text** section 8) are used to
430 ensure that selection never drives these parameters to unrealistic values.
431
432 Fifth is conversion of a TF from being an activator to being a repressor, and vice versa. The signal
433 is always an activator, and does not evolve.
434
435 Importantly, this scheme allows for divergence following gene duplication. When duplicates
436 differ due only to mutations of class 4, i.e. protein function is unchanged, we refer to them as
437 "copies" of the same gene, encoding "protein variants". Mutations in classes 2 and 5 can create
438 a new protein.
439
440

441 # Results

442 ## Functional AND-gated C1-FFLs evolve readily under selection for filtering out a short
443 ## spurious signal

444 We begin by simulating the easiest case we can devise to allow the evolution of C1-FFLs for their
445 purported function of filtering out short spurious signals. The signal is allowed to act directly on
446 the AND-gate-capable effector, so all that needs to evolve is a single activating TF between the
447 two, as well as AND-logic for the effector. We score motifs at the end of a set number of
448 generations (see Methods). Evolved C1-FFLs are scored and classified into subtypes based on
449 the presence of non-overlapping TFBSs (**Fig 3B**). The important subtype comparison for our
450 purposes being the AND-gated C1-FFL vs. the next three non-AND-gated C1-FFL types combined
451 (OR-gated, signal-controlled, and slow-TF-controlled); the remaining three logic subtypes are

vanishingly rare. The adaptive hypothesis predicts the evolution of the subtype with AND-regulatory logic, which requires both the effector to be stimulated both by the signal and by the slow TF. While all replicates show large increases in fitness, a multimodal distribution of final fitness states is observed, indicating whether or not the replicate was successful at evolving the phenotype of interest rather than becoming stuck at an alternative locally optimal phenotype (**Fig 5A**). AND-gated C1-FFLs frequently evolve in the high fitness outcomes, but not the low fitness outcomes (**Fig 5B**).

We also see C1-FFLs that, contrary to expectations, are not AND-gated; while found primarily in the low fitness replicates, some are also in the high fitness genotypes (**Fig 5B**). However, this is based on scoring motifs and their logic gates on the basis of all TFBSs, even those with two mismatches and hence low binding affinity. Unless these weak TFBSs are deleterious, they will appear quite often by chance alone. A random 8-bp sequence has probability $\binom{8}{2} \times 0.25^6 \times 0.75^2 = 0.0038$ of being a two-mismatch binding site for a given TF. In our model, a TF has the potential to recognize 137 different sites in a 150-bp cis-regulatory sequence (taking into account steric hindrance at the edges), each with 2 orientations. Thus, by chance alone a given TF will have $0.0038 \times 137 \times 2 \approx 1$ two-mismatch binding sites in a given cis-regulatory sequence (ignoring palindromes for simplicity), compared to only ~0.1 one-mismatch TFBSs. Excluding two-mismatch TFBSs when scoring motifs significantly reduces the non-AND-gated C1-FFLs, while only modestly reducing the observed frequency of adaptively evolved AND-gated C1-FFLs in the high fitness mode (**Fig 5C**).

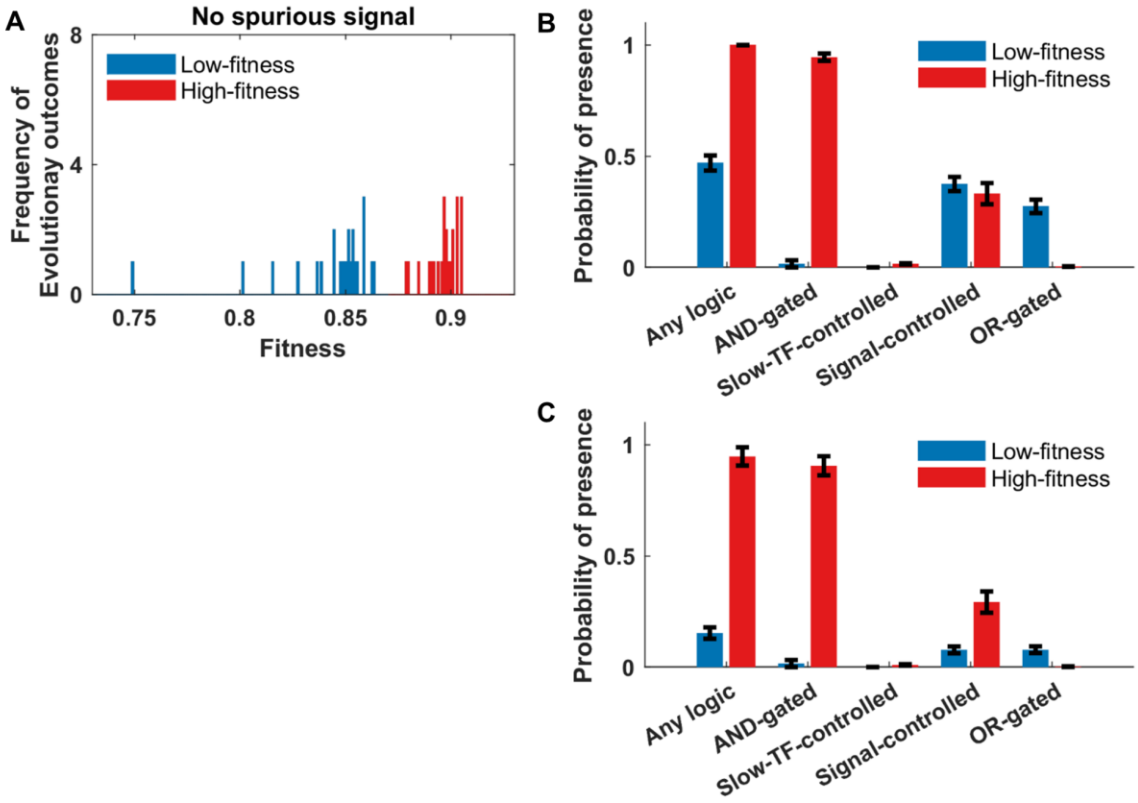


Fig 5. AND-gated C1-FFLs are associated with a successful response to selection for filtering out short spurious signals. (A) Distribution of fitness outcomes across replicate simulations, calculated as the average fitness over the last 10,000 steps of the evolutionary simulation. We divide genotypes into a low-fitness group (blue) and a high-fitness group (red) using as a threshold an observed gap in the distribution. **(B)** High fitness outcomes are characterized by the presence of an AND-gated C1-FFL. "Any logic" counts all seven subtypes shown in **Fig 3B**. Because one TRN can contain multiple C1-FFLs of different subtypes, "Any logic" will generally be less than the sum of the occurrences of all seven subtypes. See **S1 Text** section 10 for details on the calculation of the y-axis. **(C)** The over-representation of AND-gated C1-FFLs becomes even more pronounced relative to alternative logic-gating when weak (two-mismatch) TFBSs are excluded while scoring motifs. Data are shown as mean±SE of the occurrence over replicate evolution simulations. n = 23 for high-fitness group, and n = 24 for low-fitness group.

To confirm the functionality of these AND-gated C1-FFLs, we mutated the evolved genotype in two different ways (**Fig 6A**) to remove the AND regulatory logic. As expected, this lowers fitness in the presence of the short spurious signal but increases fitness in the presence of constant signal, with a net reduction in fitness (**Fig 6B**). This is consistent with AND-gated C1-FFLs representing a tradeoff, by which a more rapid response to a true signal is sacrificed in favor of the greater reliability of filtering out short spurious signals.

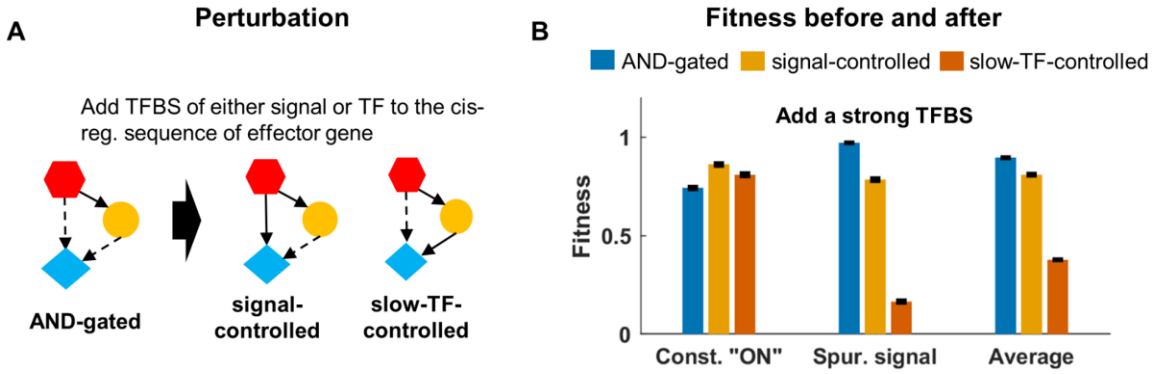
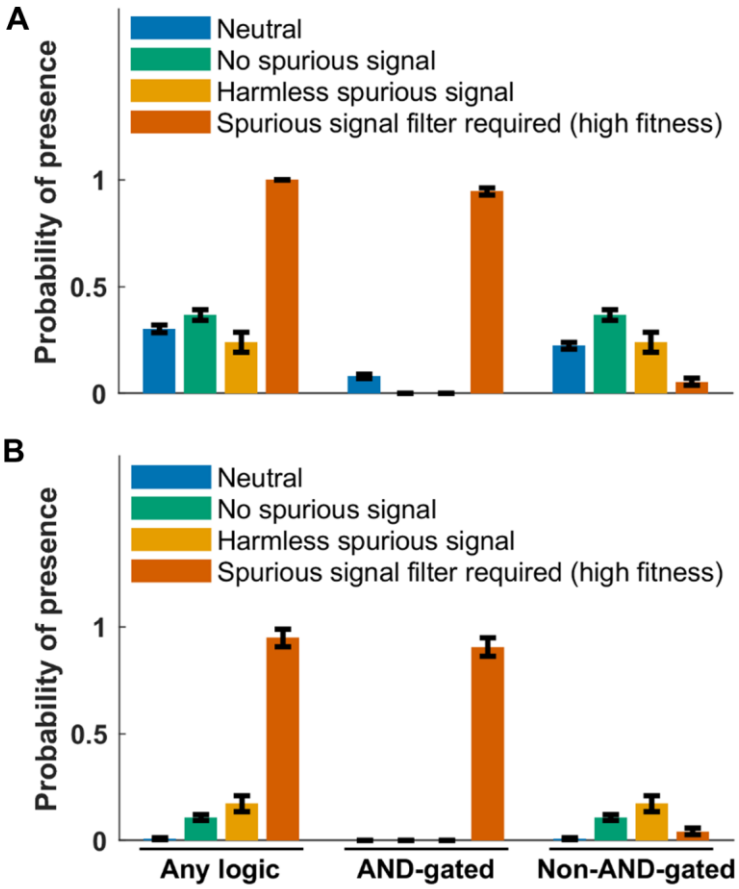


Fig 6. Destroying the AND-logic of a C1-FFL removes its ability to filter out short spurious signals. (**A**) For each of the *n* = 23 replicates in the high fitness group in **Fig 5**, we perturbed the AND-logic in two ways, by adding one binding site of either the signal or the slow TF to the cis-regulatory sequence of the effector gene, done for the subset of evolutionary steps for that replicate with AND-gated C1-FFLs and lacking other potentially confounding motifs (see **S1 Text** section 11 for details). **(B)** Destroying the AND-logic slightly increases the ability to respond to the signal, but leads to a larger loss of fitness when short spurious signals are responded to. Data are shown as mean±SE over replicate evolutionary simulations.

505 To test the extent to which C1-FFLs can evolve non-adaptively, we simulated evolution under
506 three negative control conditions: 1) neutrality, i.e. all mutations are accepted to become the
507 new resident genotype; 2) no spurious signal, i.e. the effector should be expressed under a
508 constant "ON" signal and not under a constant "OFF" signal; 3) harmless spurious signal, i.e. the
509 effector should be expressed under a constant "ON" environment whereas effector expression
510 in the "OFF" environment with short spurious signals is neither punished nor rewarded beyond
511 the cost of unnecessary gene expression. AND-gated C1-FFLs evolve much less often under all
512 three negative control conditions (**Fig 7**). Non-AND-gated C1-FFLs do evolve under the negative
513 control conditions (**Fig 7A**), but disappear when weak TFBSs are excluded during motif scoring
514 (**Fig 7B**).

515



516

Fig 7. Selection for filtering out short spurious signals is the primary cause of C1-FFLs. TRNs are evolved under different selection conditions, and we score the probability that at least one C1-FFL is present (**S1 Text** section 10). Weak (two-mismatch) TFBSs are included **(A)** or excluded **(B)** during motif scoring. Data are shown as mean±SE over evolutionary replicates. C1-FFL occurrence is similar for high-fitness and low-fitness outcomes in control selective conditions (**S3 Fig**), and so all evolutionary outcomes were combined. *n* = 30 for "Neutral", *n* = 34 for "No spurious signal", *n* = 30 for "Harmless spurious signal". "Spurious signal filter required (high fitness subset)" uses the same data as in **Fig 5**.

Diamond motifs are an alternative adaptation in more complex networks

Sometimes the source signal will not be able to directly regulate an effector, and must instead operate via a longer regulatory pathway involving intermediate TFs [61]. In this case, even if the signal itself takes the idealized form shown in **Fig 4**, its shape after propagation may become distorted by the intrinsic processes of transcription. Motifs are under selection to handle this distortion.

To enforce indirect regulation, we ran simulations in which the signal was not allowed to bind to the cis-regulatory sequence of effector genes. The fitness distribution of the evolutionary replicates has only one mode (**S4 Fig**), so we compared the highest fitness, lowest fitness, and median fitness replicates. In agreement with results when direct regulation is allowed, genotypes of low and medium fitness contain few AND-gated C1-FFLs, while high fitness genotypes contain many (**Fig 8A, left**).

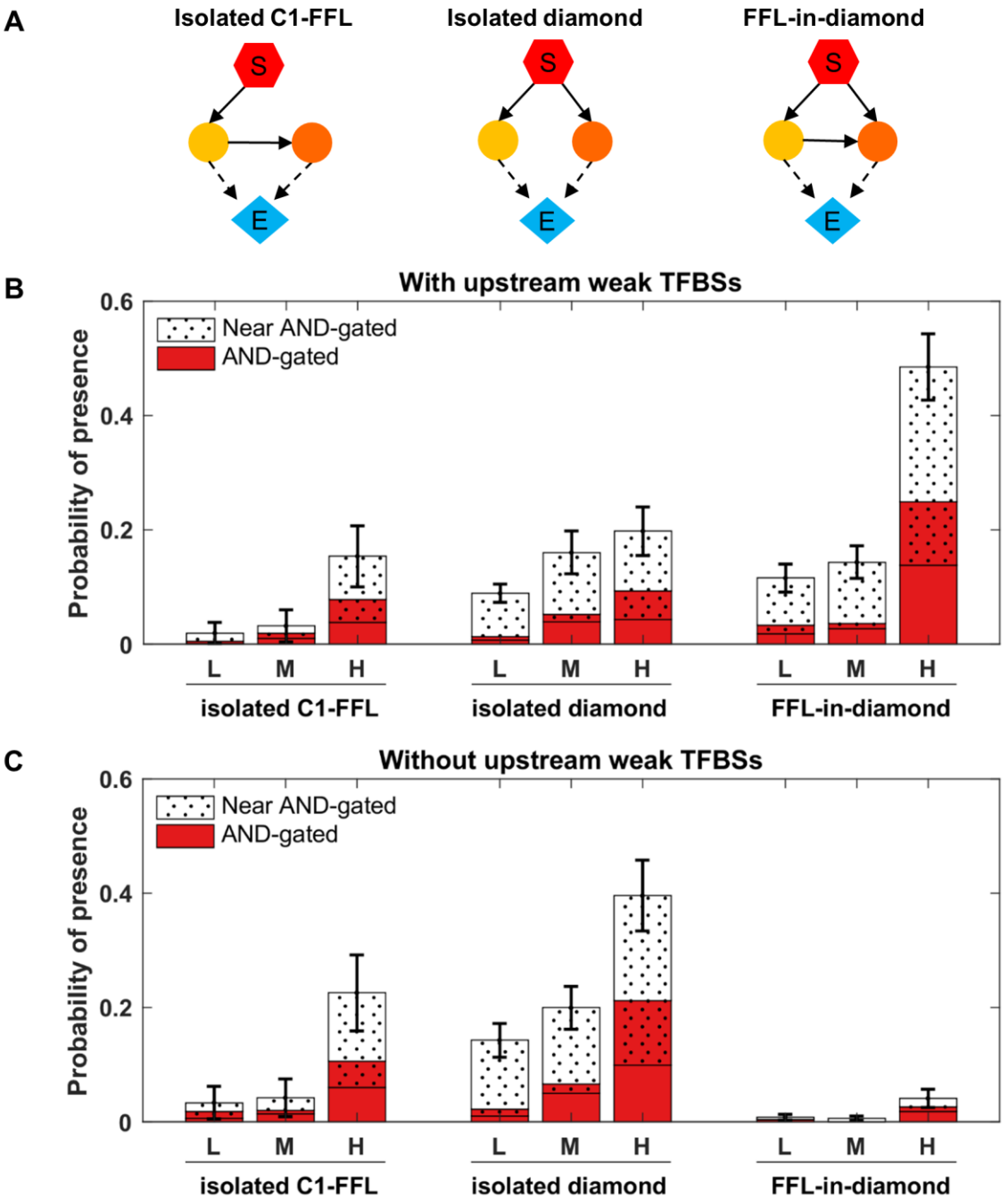
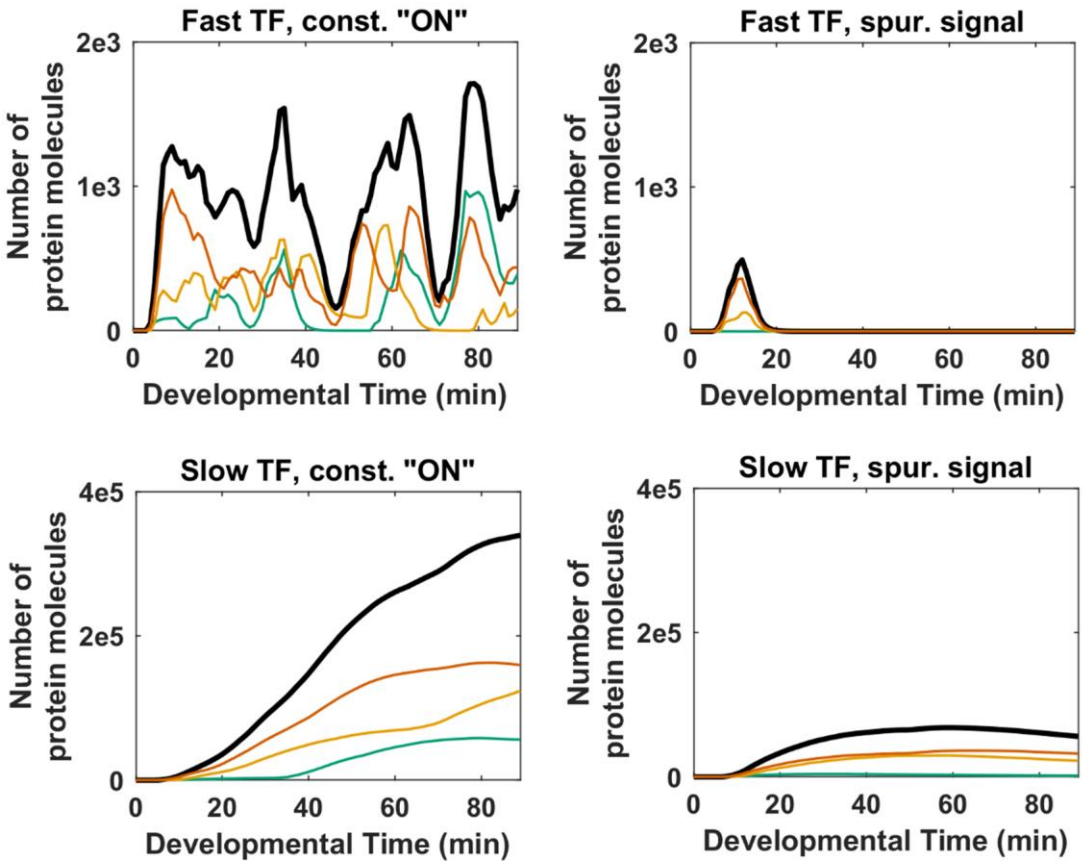


Fig 8. Both AND-gated C1-FFLs and AND-gated diamonds (A) are associated with high fitness in complex networks under selection to filter out short spurious signals. Out of 115 simulations (**S4 Fig**), we took the 30 with the highest fitness (H), the 30 with the lowest fitness (L), and 30 of around median fitness (M). AND-gated motifs are scored while including weak TFBSs, near-AND-

545 gated motifs are those scored only when these are excluded. It is possible for the same
546 genotype to contain one of each, resulting in overlap between the red AND-gated columns and
547 the dotted near-AND-gated columns. Weak TFBSs upstream, i.e. not in the effector, are shown
548 both included (**B**) and excluded (**C**). See **S1 Text** section 10 for y-axis calculation details. Error
549 bars show mean±SE over replicate evolutionary simulations.
550
551 While visually examining the network context of these C1-FFLs, we discovered that many were
552 embedded within AND-gated "diamonds" to form "FFL-in-diamonds" (**Fig 8A right**). This led us
553 to discover that AND-gated diamonds also occurred frequently without AND-gated C1-FFLs to
554 form "isolated diamonds" (**Fig 8A middle**). Note that it is in theory possible, but in practice
555 uncommon, for diamonds to be part of more complex conjugates. Systematically scoring the
556 AND-gated isolated diamond motif confirmed its high occurrence (**Fig 8B, middle**).
557
558 An AND-gated C1-FFL integrates information from a short/fast regulatory pathway with
559 information from a long/slow pathway, in order to filter out short spurious signals. A diamond
560 achieves the same end of integrating fast and slow transmitted information via differences in
561 the gene expression dynamics of the two regulatory pathways, rather than via topological length
562 (**Fig 9**).
563
564 Note that a simple transcriptional cascade, signal -> TF -> effector, has also been found
565 experimentally to filter out short spurious signals, e.g. when the intermediate TF is rapidly
566 degraded, dampening the effect of a brief signal [62]. Two such transcriptional cascades
567 involving different intermediate TFs form a diamond, so the utility of a single cascade is a
568 potential explanation for the high prevalence of double-cascade diamonds. However, in this

569 case we would have no reason to expect marked differences in expression dynamics between
570 the two TFs, as illustrated in **Fig 9**. We will also see below that AND-gates evolve between the
571 two cascades.
572



573

574 **Fig 9. The two intermediate TFs in an AND-gated "diamond" motif have different expression**
575 **dynamics and propagate the signal at different speeds.** The expression of the two TFs in one
576 representative AND-gated isolated diamond from a high-fitness genotype in **Fig 8B** is shown.
577 Each TFs is a different protein, and each is encoded by 3 gene copies, shown separately in
578 colors, with the total in thick black. The expression of one TF plateaus faster than that of the
579 other; this is characteristic of the AND-gated diamond motif, and leads to the same functionality
580 as the AND-gated C1-FFL. The two TFs are indistinguishable topologically, but can be easily and

reliably assigned identities as "fast" and "slow" by using the fact that the fast TF degrades faster (has higher $r_{protein\_deg}$). We use the geometric mean $r_{protein\_deg}$ over gene copies of a TF in order to differentiate the two TFs for analysis in **Fig 9** and elsewhere.

Weak TFBSs make motif scoring more difficult

Results depend on whether we include weak TFBSs when scoring motifs. Weak TFBSs can either be in the effector's cis-regulatory region, affecting how the regulatory logic is scored, or upstream, affecting only the presence or absence of motifs. When a motif is scored as AND-gated only when two-mismatch TFBSs in the effector are excluded, we call it a "near-AND-gated" motif. Recall from **Fig 3B** that effector expression requires two TFs to be bound, with only one TFBS of each type creating an AND-gate. When a second, two-mismatch TFBS of the same type is present, we have a near-AND-gate. TFs may bind so rarely to this weak affinity TFBS that its presence changes little, making the regulatory logic still effectively AND-gated. A near-AND-gated motif may therefore evolve for the same adaptive reasons as an AND-gated one. **Fig 8B** and **C** shows that both AND-gated and near-AND-gated motifs are enriched in the high fitness genotypes.

When we exclude upstream weak TFBSs while scoring motifs, FFL-in-diamonds are no longer found, while the occurrence of isolated C1-FFLs and diamonds increases (**Fig 8C**). This makes sense, because adding one weak TFBS, which can easily happen by chance alone, can convert an isolated diamond or C1-FFL into a FFL-in-diamond (added between intermediate TFs, or from signal to slow TF, respectively).

AND-gated isolated C1-FFLs appear mainly in the highest fitness outcomes, while AND-gated isolated diamonds appear in all fitness groups (**Fig 8C**), suggesting that diamonds are easier to evolve. 18 out of 30 high-fitness evolutionary replicates are scored as having a putatively adaptive AND-gated or near-AND-gated motif in at least 50% of their evolutionary steps when upstream weak TFBSs are ignored (close to addition of bars in **Fig 8C**, because these two AND-gated motifs rarely coexist in a high-fitness genotype). The remaining 12 have more complex arrangements of weak TFBSs that mimic a single strong one.

Just as for the AND-gated C1-FFLs evolved under direct regulation and analyzed in **Fig 6**, perturbation analysis supports an adaptive function for AND-gated C1-FFLs and diamonds evolved under indirect regulation (**Fig 10A.i, 10B.i**). Breaking the AND-gate logic of these motifs by adding a (strong) TFBS to the effector cis-regulatory region reduces the fitness under the spurious signal but increases it under the constant "ON" beneficial signal, resulting in a net decrease in the overall fitness.

If we add a two-mismatch TFBS instead, this converts an AND-gated motif to a near-AND-gated motif. This lowers fitness only when the extra link is from the slow TF to the effector, and not when the extra link is from the fast TF to the effector (**Fig 10B.ii, 10C.ii)**. Indeed, these extra links are tolerated during evolution too: if we take the 7 high-fitness replicates that contain a near-AND-gated C1-FFL in at least 5% of the evolutionary steps, in all 7 cases this motif is near-AND-gated rather than AND-gated because of an extra weak TFBS for the fast TF, while this is never due to a weak TFBS for the slow TF in C1-FFLs. Similarly, out of the 20 high-fitness replicates that contain a near-AND-gated diamond, 11 cases are primarily because of an extra weak TFBS of the fast TF, 9 cases (all of them OR-gated) are because of weak TFBSs for both TFs,

628 and no cases are primarily due to an extra TFBS for the slow TF. By chance alone, fast and slow
629 TF should be equally likely to contribute the weak TFBS that makes a motif near-AND-gated
630 rather than AND-gated. This non-random occurrence of weak TFBSs creating near-AND-gates
631 illustrates how even weak TFBSs can be shaped by selection against some (but not all) motif-
632 breaking links.

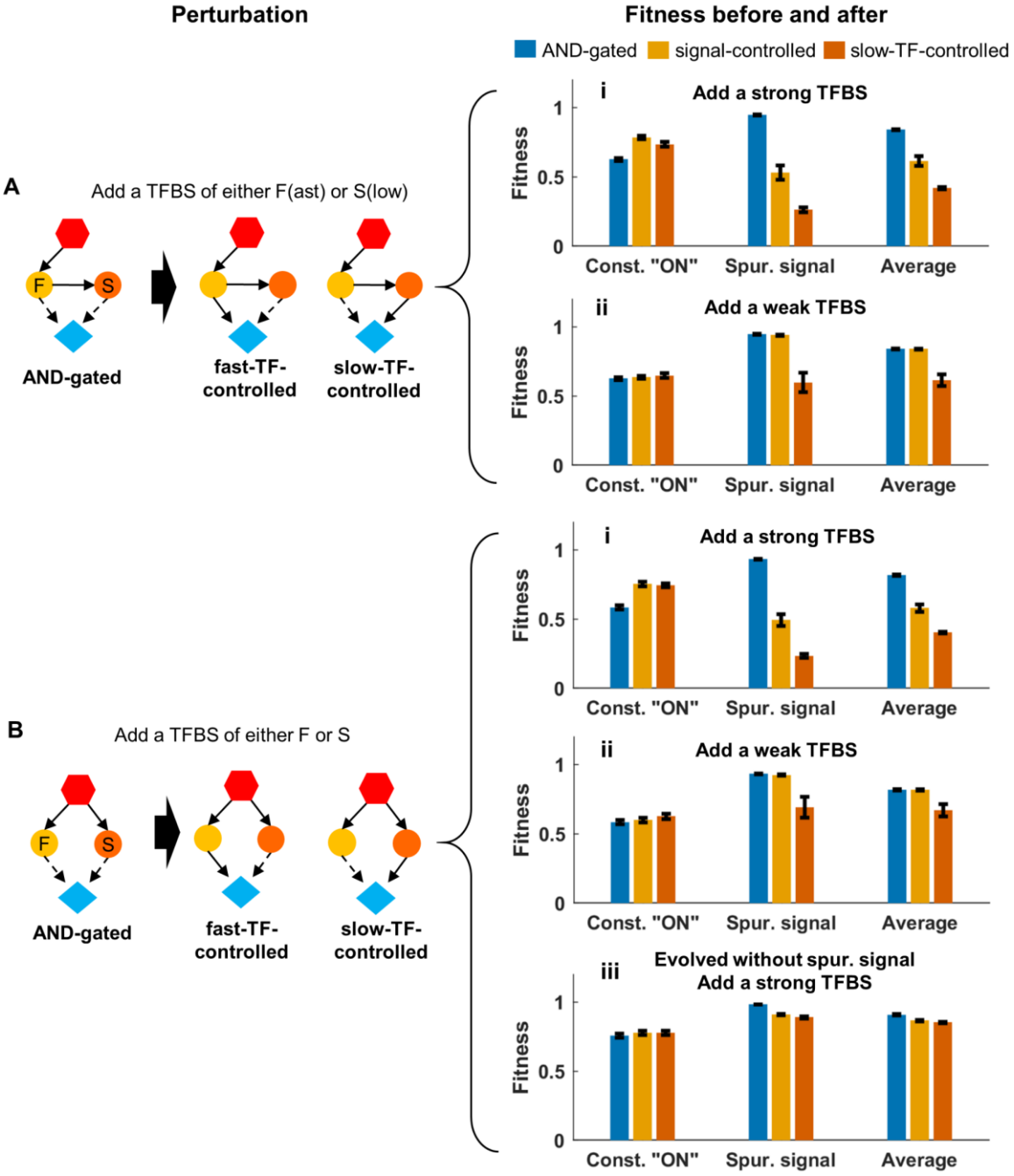
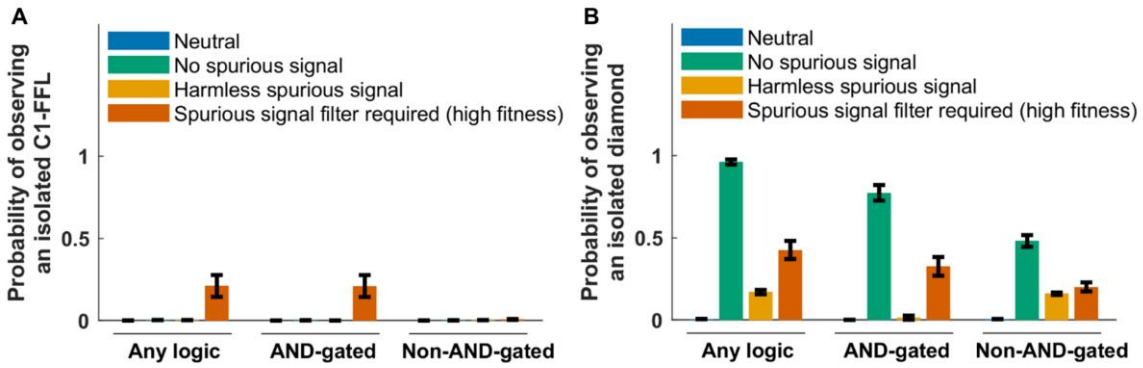


Fig 10. Perturbation analysis shows that AND-gated C1-FFLs (A) and diamonds (B) filter out short spurious signals. We add a strong TFBS **(i)** or a two-mismatch TFBS **(ii)** or **(iii)**; the latter creates near-AND-gated motifs. Allowing the effector to respond to the slow TF alone slightly increases the ability to respond to the signal, but leads to a larger loss of fitness when effector expression is undesirable. Allowing the effector to respond to the fast TF alone does so only when the conversion uses a strong TFBS not a two-mismatch TFBS. **(A)** We perform the perturbation on 5 of the 11 high-fitness replicates from **Fig 8B** that evolved an AND-gated C1-FFL. **(B)** (**i**) and (**ii**) are based on 4 of the 26 high-fitness replicates from selection to filter out short spurious external signals (**Fig 8B)**, (**iii**) is based on 18 of the 31 replicates from selection for signal recognition in the absence of an external spurious signal (**Fig 11B**). The 26 and 31 replicates were the ones with AND-gated diamond. Replicate exclusion was based on the co-occurrence of other motifs with the potential to confound results (see **S1 Text** section 11 for details). Data are shown as mean±SE of the averaged fitness over replicate evolutionary simulations.

AND-gated isolated diamonds also evolve in the absence of external spurious signals

We simulated evolution under the same three control conditions as before, this time without allowing the signal to directly regulate the effector. In the “no spurious signal” and “harmless spurious signal” control conditions, motif frequencies are similar between low and high fitness genotypes (**S5 Fig, S6 Fig)**, and so our analysis includes all evolutionary replicates. When weak (two-mismatch) TFBSs are excluded, AND-gated isolated C1-FFLs are seen only after selection for filtering out a spurious signal, and not under other selection conditions (**Fig 11A**). However, AND-gated isolated diamonds also evolve in the absence of spurious signals, indeed at even higher frequency (**Fig 11B**). Results including weak TFBSs are similar (**S7 Fig**).

657 Perturbing the AND-gate logic in these isolated diamonds reduces fitness via effects in the
658 environment where expressing the effector is deleterious (**Fig 10B.iii**). Even in the absence of
659 external short spurious signals, the stochastic expression of intermediate TFs might effectively
660 create short spurious signals when the external signal is set to "OFF". It seems that AND-gated
661 diamonds evolve to mitigate this risk, but that AND-gated C1-FFLs do not. The duration of
662 internally generated spurious signals has an exponential distribution, which means that the
663 optimal filter would be one that does not delay gene expression [63]. The two TFs in an AND-
664 gated diamond can be activated simultaneously, but they must be activated sequentially in an
665 AND-gated C1-FFL; the shorter delays possible with AND-gated diamonds might explain why
666 only diamonds and not FFLs evolve to filter out intrinsic noise in gene expression.
667



668

669 **Fig 11. Selection for filtering out a short spurious signal is the primary way to evolve AND-**
670 **gated isolated C1-FFLs (A), but AND-gated isolated diamonds also evolve in the absence of**
671 **spurious signals (B).** The selection conditions are the same as in **Fig 7**, but we do not allow the
672 signal to directly regulate the effector. When scoring motifs, we exclude all two-mismatch
673 TFBSs; more comprehensive results are shown in **S7 Fig**. See **S1 Text** section 10 for the
674 calculation of y-axis. Data are shown as mean±SE over evolutionary replicates. *n* = 30 for

675 "Neutral", *n* = 50 for "No spurious signal", and *n* = 60 for "Harmless spurious signal". We reused
676 data from **Fig 8** for "Spurious signal filter required (high fitness)", *n* = 30.
677
678

679 ## Discussion

680 There has never been sufficient evidence to satisfy evolutionary biologists that motifs in TRNs
681 represent adaptations for particular functions. Critiques by evolutionary biologists to this effect
682 [13-23] have been neglected, rather than answered, until now. While C1-FFLs can be conserved
683 across different species [64-67], this does not imply that specific "just-so" stories about their
684 function are correct. In this work, we study the evolution of AND-gated C1-FFLs, which are
685 hypothesized to be adaptations for filtering out short spurious signal [3]. Using a novel and more
686 mechanistic computational model to simulate TRN evolution, we found that AND-gated C1-FFLs
687 evolve readily under selection for filtering out a short spurious signal, and not under control
688 conditions. Our results support the adaptive hypothesis about C1-FFLs.
689
690 Previous studies have also attempted to evolve adaptive motifs in a computational TRN,
691 successfully under selection for circadian rhythm and for multiple steady states [68], and
692 unsuccessfully under selection to produce a sine wave in response to a periodic pulse [23]. Our
693 successful simulation might offer some methodological lessons, especially a focus on high-
694 fitness evolutionary replicates, which was done by us and by Burda et al. [68] but not by Knabe
695 et al. [23]. Knabe et al. [23] suggested that including a cost for gene expression may suppress
696 unnecessary links and promote motifs. However, we found AND-gated C1-FFLs still evolve in the
697 high-fitness genotypes under selection for filtering out a spurious signal, even when there is no
698 cost of gene expression (**S8 Fig**).

699

700 AND-gated C1-FFLs express an effector after a noise-filtering delay when the signal is turned on,
701 but shut down expression immediately when the signal is turned off, giving rise to a "sign-
702 sensitive delay" [3, 7]. Rapidly switching off has been hypothesized to be part of their selective
703 advantage, above and beyond the function of filtering out short spurious signals [63]. We
704 selected only for filtering out a short spurious signal, and not for fast turn-off, and found that
705 this was sufficient for the adaptive evolution of AND-gated C1-FFLs.

706

707 Most previous research on C1-FFLs has used an idealized implementation (e.g. a square wave) of
708 what a short spurious signal entails [4, 63, 69]. In real networks, noise arises intrinsically in a
709 greater diversity of forms, which our model does more to capture. Even when a "clean" form of
710 noise enters a TRN, it subsequently gets distorted with the addition of intrinsic noise [70].
711 Intrinsic noise is ubiquitous and dealing with it is an omnipresent challenge for selection.
712 Indeed, we see adaptive diamonds evolve to suppress intrinsic noise, even when we select in
713 the complete absence of extrinsic spurious signals.

714

715 Our model, while complex for a model and hence capable of capturing intrinsic noise, is
716 inevitably less complex than the biological reality. However, we hope to have captured key
717 phenomena, albeit in simplified form. E.g., a key phenomenon is that TFBSs are not simply
718 present vs. absent but can be strong or weak, i.e. the TRN is not just a directed graph, but its
719 connections vary in strength. Our model, like that of Burda et al. [68] in the context of circadian
720 rhythms, captures this fact by basing TF binding affinity on the number of mismatch deviations
721 from a consensus TFBS sequence. While in reality, the strength of TF binding is determined by
722 additional factors, such as broader nucleic context and cooperative behavior between TFs

723 (reviewed in Inukai et al. [71]), these complications are unlikely to change the basic dynamics of
724 frequent appearance of weak TFBSs and enhanced mutational accessibility of strong TFBSs from
725 weak ones. Similarly, AND-gating can be quantitative rather than qualitative [72], a
726 phenomenon that weak TFBSs in our model provide a simplified version of. Note that our
727 model, while powerful in some ways, is computationally limited to small TRNs.
728 Core links in adaptive motifs involve strong not weak TFBSs. However, weak (two-mismatch)
729 TFBSs can create additional links that prevent an adaptive motif from being scored as such.
730 Some potential additional links are neutral while others are deleterious; the observed links are
731 thus shaped by this selective filter, without being adaptive. Note that there have been
732 experimental reports that even weak TFBSs can be functionally important [73, 74]; these might,
733 however, better correspond to 1-mismatch TFBSs in our model than two-mismatch TFBSs.
734 Ramos et al. [74] and Crocker et al. [73] identified their “weak” TFBSs in comparison to the
735 strongest possible TFBS, not in comparison to the weakest still showing affinity above baseline.
736
737 A striking and unexpected finding of our study was that AND-gated diamonds evolved as an
738 alternative motif for filtering out short spurious external signals, and that these, unlike FFLs,
739 were also effective at filtering out intrinsic noise. Diamonds are not overrepresented in the TRNs
740 of bacteria [2] or yeast [75], but are overrepresented in signaling networks (in which post-
741 translational modification plays a larger role) [76], and in neuron networks [1]. In our model, we
742 treated the external signal as though it were a transcription factor, simply as a matter of
743 modeling convenience. In reality, signals external to a TRN are by definition not TFs (although
744 they might be modifiers of TFs). This means that our indirect regulation case, in which the signal
745 is not allowed to directly turn on the effector, is the most appropriate one to analyze if our
746 interest is in TRN motifs that mediate contact between the two. Note that if we were to score

747 the signal as not itself a TF, we would observe adaptive C1-FFLs but not diamonds in this case, in
748 agreement with the TRN data. However, this TRN data might miss functional diamond motifs
749 that spanned levels of regulatory organization, i.e. that included both transcriptional and other
750 forms of regulation. The greatest chance of finding diamonds within TRNs alone come from
751 complex and multi-layered developmental cascades, rather than bacterial or yeast [77]. Multiple
752 interwoven diamonds are hypothesized to be embedded with multi-layer perceptrons that are
753 adaptations for complex computation in signaling networks [30].
754
755 The function of a motif relies ultimately on its dynamic behavior, with topology merely a means
756 to that end. The C1-FFL motif is based on two pathways between signal and effector, one much
757 faster than the other, which is achieved by making them different lengths. This same function
758 was achieved non-topologically in our adaptively evolved diamond motifs. Multiple motifs have
759 previously been found capable of generating the same steady state expression pattern [21];
760 here we find multiple motifs for a much more complex function.
761
762 It is difficult to distinguish adaptations from “spandrels” [8]. Standard procedure is to look for
763 motifs that are more frequent than expected from some randomized version of a TRN [2, 78].
764 For this method to work, this randomization must control for all confounding factors that are
765 non-adaptive with respect to the function in question, from patterns of mutation to a general
766 tendency to hierarchy – a near-impossible task. Our approach to a null model is not to
767 randomize, but to evolve with and without selection for the specific function of interest. This
768 meets the standard of evolutionary biology for inferring the adaptive nature of a motif [13-23].
769
770

Acknowledgements

Work was supported by the University of Arizona and by a Pew Scholarship to JM, John Templeton Foundation grant 39667 to JM and KX, and by National Institutes of Health grants R35GM118170 to MLS and R01GM076041 to JM and AKL. We thank Hinrich Boeger for helpful discussions and careful reading of the manuscript, Jasmin Uribe for early work on this project, and the high-performance computing center at the University of Arizona for generous allocations.

References

1. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. Science. 2002;298:824-827.

2. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet. 2002;31:64-68. doi: 10.1038/ng881.

3. Alon U. Network motifs: theory and experimental approaches. Nature Reviews Genetics. 2007;8:450-461.

4. Mangan S, Alon U. Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci USA. 2003;100:11980-11985. doi: 10.1073/pnas.2133841100.

5. Jaeger KE, Pullen N, Lamzin S, Morris RJ, Wigge PA. Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. The Plant Cell. 2013;25:820-833. doi: 10.1105/tpc.113.109355.

6. Peter IS, Davidson EH. Assessing regulatory information in developmental gene regulatory networks. Proc Natl Acad Sci USA. 2017;114:5862-5869. doi: 10.1073/pnas.1610616114.

7. Mangan S, Zaslaver A, Alon U. The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol. 2003;334:197-204. doi: 10.1016/j.jmb.2003.09.049.

8. Gould SJ, Lewontin RC. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. Proc R Soc Lond, Ser B: Biol Sci. 1979;205:581-598.

9. Graur D, Zheng Y, Price N, Azevedo RBR, Zufall RA, Elhaik E. On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free

802 Gospel of ENCODE. Genome Biology and Evolution. 2013;5:578-590. doi:
803 10.1093/gbe/evt028.
804 10. Masel J, Promislow DEL. Answering evolutionary questions: A guide for mechanistic
805 biologists. Bioessays. 2016;38:704-711. doi: 10.1002/bies.201600029.
806 11. Widder S, Solé R, Macía J. Evolvability of feed-forward loop architecture biases its
807 abundance in transcription networks. BMC Systems Biology. 2012;6:7. doi:
808 10.1186/1752-0509-6-7.
809 12. Cordero OX, Hogeweg P. Feed-forward loop circuits as a side effect of genome
810 evolution. Mol Biol Evol. 2006;23:1931-1936.
811 13. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L. Comment on "Network Motifs:
812 Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and
813 Designed Networks". Science. 2004;305:1107. doi: 10.1126/science.1099334.
814 14. Jenkins D, Stekel D. De Novo Evolution of Complex, Global and Hierarchical Gene
815 Regulatory Mechanisms. J Mol Evol. 2010;71:128-140.
816 15. Lynch M. The evolution of genetic networks by non-adaptive processes. Nature Reviews
817 Genetics. 2007;8:803-813. doi: 10.1038/nrg2192.
818 16. Mazurie A, Bottani S, Vergassola M. An evolutionary and functional assessment of
819 regulatory network motifs. Genome Biol. 2005;6:R35.
820 17. Solé RV, Valverde S. Are network motifs the spandrels of cellular complexity? Trends
821 Ecol Evol. 2006;21:419-422.
822 18. Tsuda ME, Kawata M. Evolution of Gene Regulatory Networks by Fluctuating Selection
823 and Intrinsic Constraints. PLoS Comp Biol. 2010;6:e1000873.
824 19. Wagner A. Does Selection Mold Molecular Networks? Sci STKE. 2003;2003:pe41. doi:
825 10.1126/stke.2003.202.pe41.

826 20. Kuo PD, Banzhaf W, Leier A. Network topology and the evolution of dynamics in an
827 artificial genetic regulatory network model created by whole genome duplication and
828 divergence. BioSyst. 2006;85:177-200. doi: 10.1016/j.biosystems.2006.01.004.
829 21. Payne JL, Wagner A. Function does not follow form in gene regulatory circuits. Scientific
830 Reports. 2015;5:13015. doi: 10.1038/srep13015.
831 22. Ruths T, Nakhleh L. Neutral forces acting on intragenomic variability shape the
832 *Escherichia coli* regulatory network topology. Proc Natl Acad Sci USA. 2013;110:7754-
833 7759. doi: 10.1073/pnas.1217630110.
834 23. Knabe JF, Nehaniv CL, Schilstra MJ. Do motifs reflect evolved function?—No convergent
835 evolution of genetic regulatory network subgraph topologies. BioSyst. 2008;94:68-74.
836 doi: 10.1016/j.biosystems.2008.05.012.
837 24. Kærn M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories
838 to phenotypes. Nature Reviews Genetics. 2005;6:451-464. doi: 10.1038/nrg1615.
839 25. Raser JM, O'Shea EK. Noise in Gene Expression: Origins, Consequences, and Control.
840 Science. 2005;309:2010-2013.
841 26. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010;467:167–
842 173. doi: 10.1038/nature09326.
843 27. Draghi J, Whitlock M. Robustness to noise in gene expression evolves despite epistatic
844 constraints in a model of gene networks. Evolution. 2015;69:2345-2358. doi:
845 10.1111/evo.12732.
846 28. Jenkins DJ, Stekel DJ. A New Model for Investigating the Evolution of Transcription
847 Control Networks. Artif Life. 2009;15:259-291. doi: doi:10.1162/artl.2009.Stekel.006.

848 29. Henry A, Hemery M, François P. ϕ-evo: A program to evolve phenotypic models of
849 biological networks. PLoS Comp Biol. 2018;14:e1006244. doi:
850 10.1371/journal.pcbi.1006244.
851 30. Alon U. An Introduction to Systems Biology: Design Principles of Biological Circuits.
852 London: Chapman and Hall/CRC; 2007.
853 31. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, et al. Genome-scale
854 identification of nucleosome positions in *S. cerevisiae*. Science. 2005;309:626-630.
855 32. Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of
856 binding motifs. Trends Genet. 2009;25:434-440. doi: 10.1016/j.tig.2009.08.003.
857 33. Maerkl SJ, Quake SR. A Systems Approach to Measuring the Binding Energy Landscapes
858 of Transcription Factors. Science. 2007;315:233-237. doi: 10.1126/science.1131007.
859 34. Zhu J, Zhang MQ. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*.
860 Bioinformatics. 1999;15:607-611. doi: 10.1093/bioinformatics/15.7.607.
861 35. Sharon E, Kalma Y, Sharp A, Raveh-sadka T, Levo M, Zeevi D, et al. Articles Inferring gene
862 regulatory logic from high-throughput measurements of thousands of systematically
863 designed promoters. Nat Biotechnol. 2012;30:521-530. doi: 10.1038/nbt.2205.
864 36. Benos PV, Bulyk ML, Stormo GD. Additivity in protein-DNA interactions: how good an
865 approximation is it? Nucleic Acids Res. 2002;30:4442-4451.
866 37. Park S, Chung S, Kim KM, Jung KC, Park C, Hahm ER, et al. Determination of binding
867 constant of transcription factor myc-max/max-max and E-box DNA: The effect of
868 inhibitors on the binding. Biochim Biophys Acta Gen Subj. 2004;1670:217-228. doi:
869 10.1016/j.bbagen.2003.12.007.

870 38. Nalefski EA, Nebelitsky E, Lloyd JA, Gullans SR. Single-molecule detection of transcription
871 factor binding to DNA in real time: Specificity, equilibrium, and kinetic parameters.
872 Biochemistry. 2006;45:13794-13806. doi: 10.1021/bi0602011.
873 39. Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, et al. A high-resolution atlas of
874 nucleosome occupancy in yeast. Nat Genet. 2007;39:1235-44.
875 40. Brown CR, Mao C, Falkovskaia E, Jurica MS, Boeger H. Linking Stochastic Fluctuations in
876 Chromatin Structure and Gene Expression. PLoS Biol. 2013;11:e1001621. doi:
877 10.1371/journal.pbio.1001621.
878 41. Mao C, Brown CR, Falkovskaia E, Dong S, Hrabeta-Robinson E, Wenger L, et al.
879 Quantitative analysis of the transcription control mechanism. Mol Syst Biol. 2010;6:431.
880 doi: 10.1038/msb.2010.83.
881 42. Shahbazian MD, Grunstein M. Functions of Site-Specific Histone Acetylation and
882 Deacetylation. Annu Rev Biochem. 2007;76:75-100. doi:
883 10.1146/annurev.biochem.76.052705.162114.
884 43. Voss TC, Hager GL. Dynamic regulation of transcriptional states by chromatin and
885 transcription factors. Nature Reviews Genetics. 2013;15:69-81. doi: 10.1038/nrg3623.
886 44. Katan-Khaykovich Y, Struhl K. Dynamics of global histone acetylation and deacetylation
887 in vivo: rapid restoration of normal histone acetylation status upon removal of
888 activators and repressors. Genes Dev. 2002;16:743-52. doi: 10.1101/gad.967302.
889 45. Courey AJ, Jia S. Transcriptional repression: the long and the short of it. Genes Dev.
890 2001;15:2786-2796. doi: 10.1101/gad.939601.and.
891 46. Poss ZC, Ebmeier CC, Taatjes DJ. The Mediator complex and transcription regulation. Crit
892 Rev Biochem Mol Biol. 2013;48:575-608. doi: 10.3109/10409238.2013.840259.

47. Decker KB, Hinton DM. Transcription Regulation at the Core: Similarities Among Bacterial, Archaeal, and Eukaryotic RNA Polymerases. Annu Rev Microbiol. 2013;67:113-139. doi: 10.1146/annurev-micro-092412-155756.

48. Roy AL, Singer DS. Core promoters in transcription: old problem, new insights. Trends Biochem Sci. 2015;40:165-171. doi: 10.1016/j.tibs.2015.01.007.

49. Pelechano V, Chávez S, Pérez-Ortín JE. A Complete Set of Nascent Transcription Rates for Yeast Genes. PLoS ONE. 2010;5:e115560.

50. Guillemette B, Bataille AR, Gevry N, Adam M, Blanchette M, Robert F, et al. Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol. 2005;3:e384.

51. Dujon B. The yeast genome project: what did we learn? Trends Genet. 1996;12:263-270. doi: 10.1016/0168-9525(96)10027-5.

52. Niño CA, Hérissant L, Babour A, Dargemont C. mRNA nuclear export in yeast. Chemical Reviews. 2013;113:8523-8545. doi: 10.1021/cr400002g.

53. Smith C, Lari A, Derrer CP, Ouwehand A, Rossouw A, Huisman M, et al. In vivo single-particle imaging of nuclear mRNA export in budding yeast demonstrates an essential role for Mex67p. J Cell Biol. 2015;211:1121-1130. doi: 10.1083/jcb.201503135.

54. Mor A, Suliman S, Ben-Yishay R, Yunger S, Brody Y, Shav-Tal Y. Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. Nat Cell Biol. 2010;12:543-552. doi: 10.1038/ncb2056.

55. Siwiak M, Zielenkiewicz P, Jacobson A, Grebogi C, Kito K. A Comprehensive, Quantitative, and Genome-Wide Model of Translation. PLoS Comp Biol. 2010;6:e1000865. doi: 10.1371/journal.pcbi.1000865.

916 56. van Drogen F, Stucke VM, Jorritsma G, Peter M. MAP kinase dynamics in response to
917 pheromones in budding yeast. Nat Cell Biol. 2001;3:1051.
918 57. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. Journal of
919 Physical Chemistry. 1977;81:2340-2361.
920 58. Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, et al. Global
921 analysis of protein expression in yeast. Nature. 2003;425:737-741. doi:
922 10.1038/nature02046.
923 59. Kafri M, Metzl-Raz E, Jona G, Barkai N. The Cost of Protein Production. Cell Reports.
924 2016;14:22-31. doi: 10.1016/j.celrep.2015.12.015.
925 60. Kimura M. On the probability of fixation of mutant genes in a population. Genetics.
926 1962;47:713-719.
927 61. Balázsi G, Barabási AL, Oltvai ZN. Topological units of environmental signal processing in
928 the transcriptional regulatory network of *Escherichia coli*. Proc Natl Acad Sci USA.
929 2005;102:7841-7846.
930 62. Hooshangi S, Thiberge S, Weiss R. Ultrasensitivity and noise propagation in a synthetic
931 transcriptional cascade. Proc Natl Acad Sci USA. 2005;102:3581-3586. doi:
932 10.1073/pnas.0408507102.
933 63. Dekel E, Mangan S, Alon U. Environmental selection of the feed-forward loop circuit in
934 gene-regulation networks. Phys Biol. 2005;2:81.
935 64. Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, et al. Comparative analysis of
936 regulatory information and circuits across distant species. Nature. 2014;512:453-456.
937 doi: 10.1038/nature13668.

938 65. Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, et al.
939 Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of
940 Gene-Specific Repressors. Cell. 2014;157:740-752. doi: 10.1016/j.cell.2014.02.054.
941 66. Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, et al.
942 Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature.
943 2014;515:365-370. doi: 10.1038/nature13972.
944 67. Madan Babu M, Teichmann SA, Aravind L. Evolutionary Dynamics of Prokaryotic
945 Transcriptional Regulatory Networks. J Mol Biol. 2006;358:614-633. doi:
946 10.1016/j.jmb.2006.02.019.
947 68. Burda Z, Krzywicki A, Martin OC, Zagorski M. Motifs emerge from function in model
948 gene regulatory networks. Proc Natl Acad Sci USA. 2011;108:17263-17268. doi:
949 10.1073/pnas.1109435108.
950 69. Hayot F, Jayaprakash C. A feedforward loop motif in transcriptional regulation: induction
951 and repression. J Theor Biol. 2005;234:133-143.
952 70. Pedraza JM, van Oudenaarden A. Noise Propagation in Gene Networks. Science.
953 2005;307:1965-1969. doi: 10.1126/science.1109090.
954 71. Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site
955 motifs. Curr Opin Genet Dev. 2017;43:110-119. doi: 10.1016/j.gde.2017.02.007.
956 72. Wang D, Yan K-K, Sisu C, Cheng C, Rozowsky J, Meyerson W, et al. Loregic: A Method to
957 Characterize the Cooperative Logic of Regulatory Factors. PLoS Comp Biol.
958 2015;11:e1004132. doi: 10.1371/journal.pcbi.1004132.
959 73. Crocker J, Abe N, Rinaldi L, McGregor Alistair P, Frankel N, Wang S, et al. Low Affinity
960 Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. Cell.
961 2015;160:191-203. doi: 10.1016/J.CELL.2014.11.041.

962 74. Ramos AI, Barolo S. Low-affinity transcription factor binding sites shape morphogen
963 responses and enhancer evolution. Philosophical transactions of the Royal Society of
964 London Series B, Biological sciences. 2013;368:20130018. doi: 10.1098/rstb.2013.0018.
965 75. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional
966 regulatory networks in *Saccharomyces cerevisiae*. Science. 2002;298:799-804.
967 76. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, et al. Formation
968 of regulatory patterns during signal propagation in a mammalian cellular network.
969 Science. 2005;309:1078-1083. doi: 10.1126/science.1108876.
970 77. Rosenfeld N, Alon U. Response Delays and the Structure of Transcription Networks. J
971 Mol Biol. 2003;329:645-654. doi: 10.1016/S0022-2836(03)00506-0.
972 78. Kashtan N, Itzkovitz S, Milo R, Alon U. Efficient sampling algorithm for estimating
973 subgraph concentrations and detecting network motifs. Bioinformatics. 2004;20:1746-
974 1758. doi: 10.1093/bioinformatics/bth163.
975 79. SGD Project. [cited 2018 April 2]. Available from: https://yeastmine.yeastgenome.org.
976 80. Hocine S, Raymond P, Zenklusen D, Chao JA, Singer RH. Single-molecule analysis of gene
977 expression using two-color RNA labeling in live yeast. Nat Methods. 2013;10:119-121.
978 doi: 10.1038/nmeth.2305.
979 81. Larson DR, Zenklusen D, Wu B, Chao JA, Singer RH. Real-time observation of
980 transcription initiation and elongation on an endogenous yeast gene. Science.
981 2011;332:475-478. doi: 10.1126/science.1202142.
982 82. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. Precision and functional
983 specificity in mRNA decay. Proc Natl Acad Sci USA. 2002;99:5860-5865. doi:
984 10.1073/pnas.092538799.

985 83. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK. Quantification of protein half-lives in
986 the budding yeast proteome. Proc Natl Acad Sci USA. 2006;103:13004-13009. doi:
987 10.1073/pnas.0605420103.

988

Supporting information

S1 Fig. Examples of evolved phenotypes under selection for filtering out a short spurious signal. The figure shows the average expression of the effector protein over 200 replicate developmental simulations in each of the two environments. A high-fitness phenotype and a low-fitness phenotype, as defined in **Fig 5**, are shown for comparison. The signal is allowed to directly regulate the effector in these simulations.

S2 Fig. Representative fitness trajectories under selection to filter out short spurious signals. **(A)** The signal is allowed to directly regulate the effector genes. **(B)** The signal cannot directly regulate the effector genes. Note the average is weighted, with environment 2 being considered twice as common as environment 1.

S3 Fig. Genotypes evolved under control selective conditions: (A) "harmless spurious signal", and (B) "no spurious signal". There is no clear evidence of a multimodal distribution of fitness outcomes among replicates (left), and C1-FFLs occur equally in the 10 genotypes of the highest fitness vs. the 10 genotypes of the lowest fitness (right), and so the entire distribution (left) was used to produce **Fig 7**. Data are shown as mean±SE over evolutionary replicates.

S4 Fig. Fitness distrbution of 115 evolutionary replicates under selection for filtering out short spurious signals, when the signal cannot directly regulate the effector. The fitness of a replicate is the average genotype fitness over the last 10,000 evolutionary steps. Colors indicate replicates analyzed elsewhere.

S5 Fig. Evolution when responding to a spurious signal is harmless, when the signal is not allowed to directly regulate the effector. (A) Fitness distribution of 60 replicate simulations. The occurrence of both **(B)** FFL-in-diamonds and **(C)** isolated diamonds were similar in the 10 genotypes with the highest fitness vs. in 10 genotypes with the lowest fitness. Weak (two-mismatch) TFBSs are included when scoring motifs. Data are shown as mean±SE over replicates. Isolated C1-FFLs rarely evolve under this condition, therefore their occurrence is not plotted.

S6 Fig. Evolution when there is no spurious signal, when the signal is not allowed to directly regulate the effector. (A) Fitness distribution of 50 replicate simulations. The occurrence of both **(B)** FFL-in-diamonds and **(C)** isolated diamonds were similar in the 10 genotypes with the highest fitness vs. in the 10 genotypes with the lowest fitness. Weak (two-mismatch) TFBSs are included when scoring motifs. Data are shown as mean±SE over replicates. Isolated C1-FFLs rarely evolve under this condition, therefore their occurrence is not plotted.

S7 Fig. Selection for filtering out a short spurious signal is the primary way to evolve AND-gated C1-FFLs (A), but AND-gated isolated diamonds also evolve in the absence of spurious signals (B). The signal is not allowed to directly regulate the effector, and the right hand sides of **(A)** and **(B)** are identical to **Fig 11**. When scoring motifs, we either include (left) or exclude (right) all two-mismatch TFBSs in the cis-regulatory sequences of intermediate TF genes and effector genes. See **S1 Text** section 10 for the calculation of y-axis. Data are shown as mean±SE over evolutionary replicates.

S8 Fig. After removing the cost of gene expression, AND-gated C1-FFLs are still associated with a successful response to selection for filtering out a short spurious signal. The signal can

directly regulate the effector genes. **(A)** Distribution of fitness outcomes across 46 replicate simulations. **(B)** 10 out of 13 replicates with the highest fitness [the 13 replicates are in red in **(A)**] still evolve AND-gated C1-FFLs. Replicates with the 4^{th}, 6^{th}, and 8^{th} highest fitness evolve the motif shown in **(C)** rather than AND-gated C1-FFLs. The "high-fitness" group therefore replace the three replicates with replicates with the 11^{th} to 13^{th} highest fitness. Bars are mean±SE of the occurrence ove replicate evolutionary simulations. 5 replicates [blue in **(A)**] with the lowest fitness do not contain AND-gated C1-FFLs or the motif in **(C)**. **(C)** AND-gated C1-FFLs with a long arm. Note that both S and B need to be present to induce the expression of E, therefore this motif can also act as spurious signal filter.

S1 Text. Additional details of the model and algorithms