

## Research



**Cite this article:** Stewart AJ, Plotkin JB. 2013  
The evolution of complex gene regulation by  
low-specificity binding sites. *Proc R Soc B* 280:  
20131313.  
<http://dx.doi.org/10.1098/rspb.2013.1313>

Received: 30 May 2013

Accepted: 17 July 2013

### Subject Areas:

evolution, theoretical biology

### Keywords:

gene regulation, population genetics,  
*cis*-regulatory evolution

### Author for correspondence:

Joshua B. Plotkin

e-mail: [jplotkin@sas.upenn.edu](mailto:jplotkin@sas.upenn.edu)

Electronic supplementary material is available  
at <http://dx.doi.org/10.1098/rspb.2013.1313> or  
via <http://rspb.royalsocietypublishing.org>.

# The evolution of complex gene regulation by low-specificity binding sites

Alexander J. Stewart and Joshua B. Plotkin

Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

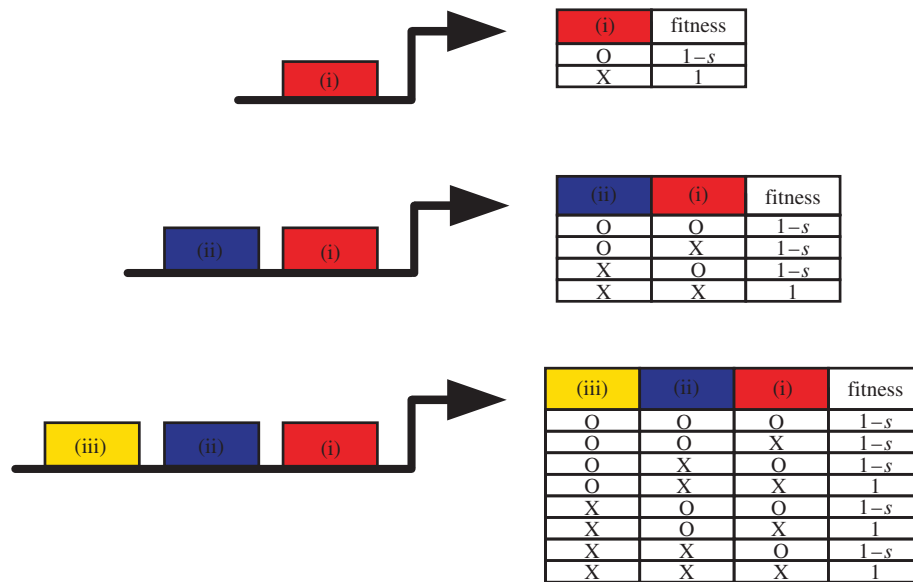
Requirements for gene regulation vary widely both within and among species. Some genes are constitutively expressed, whereas other genes require complex regulatory control. Transcriptional regulation is often controlled by a module of multiple transcription factor binding sites that, in combination, mediate the expression of a target gene. Here, we study how such regulatory modules evolve in response to natural selection. Using a population-genetic model, we show that complex regulatory modules which contain a larger number of binding sites must employ binding motifs that are less specific, on average, compared with smaller regulatory modules. This effect is extremely general, and it holds regardless of the selected binding logic that a module experiences. We attribute this phenomenon to the inability of stabilizing selection to maintain highly specific sites in large regulatory modules. Our analysis helps to explain broad empirical trends in the *Saccharomyces cerevisiae* regulatory network: those genes with a greater number of distinct transcriptional regulators feature less-specific binding motifs, compared with genes with fewer regulators. Our results also help to explain empirical trends in module size and motif specificity across species, ranging from prokaryotes to single-cellular and multi-cellular eukaryotes.

## 1. Introduction

Transcriptional regulators integrate signals from genes and the environment to ensure that the correct patterns of gene expression are maintained in the cell [1–8]. This can be a complicated task, particularly in higher eukaryotes where processes such as cell differentiation and complex inter-cellular signalling occur [1,4]. Generally, the more complex a signal integration task, the more complex the pattern of gene regulation required. On the face of it, we might expect more complex gene regulation to be carried out by binding sites of higher specificity—just as we expect a complicated machine to use high-precision components. Here we show that, in fact, the opposite is true: natural selection favours less-specific binding sites in more complex regulatory modules.

We use an established biophysical model of transcription factor binding to describe a regulatory module—namely, a set of multiple transcription factor binding sites, with a range of specificities, that in combination regulate the expression of a given target gene. For selection of a given strength on the target gene's expression level, we determine the average information of binding motifs that participate in a module of a given size, across all possible binding logics. Our analysis predicts that the binding sites maintained in a large module will each have lower information content, on average, than those maintained in a small module. We equate the size of a module with the complexity of its regulatory capacity, because larger modules can assume a greater number of distinct states and execute more complex binding logics. Thus, our analysis predicts more complex regulation by less-specific binding motifs.

This simple but counterintuitive result helps to explain two broad empirical patterns in transcription networks within and between species. First, regulatory complexity in eukaryotes is greater than in prokaryotes [9,10], and this difference is accompanied by a tendency towards less-informative binding motifs for eukaryotic transcription factors [11,12]. Second, within the yeast transcription network, we observe that those genes whose expression varies across environmental conditions, and therefore require more complex integration of



**Figure 1.** A transcriptional module is a set of distinct transcription factor binding sites that, in combination, regulate the expression of a given target gene. Module size is defined as the number of distinct binding sites (blocks (i), (ii) and (iii)). As module size increases, so too does regulatory complexity, as quantified by the number of distinct states that the module can occupy. This is illustrated using truth tables (right) with 'X' indicating a bound site and 'O' indicating an unbound site. The fitness of different binding combinations is shown in the final column. Selection favours some patterns of binding (fitness 1) and disfavours others (fitness  $1 - s$ ) and thereby determines the evolution of the binding sites in the module. (Online version in colour.)

environmental signals, tend to have more transcriptional regulators with binding motifs that each have lower specificity, compared with genes expressed consistently across environmental conditions [13]

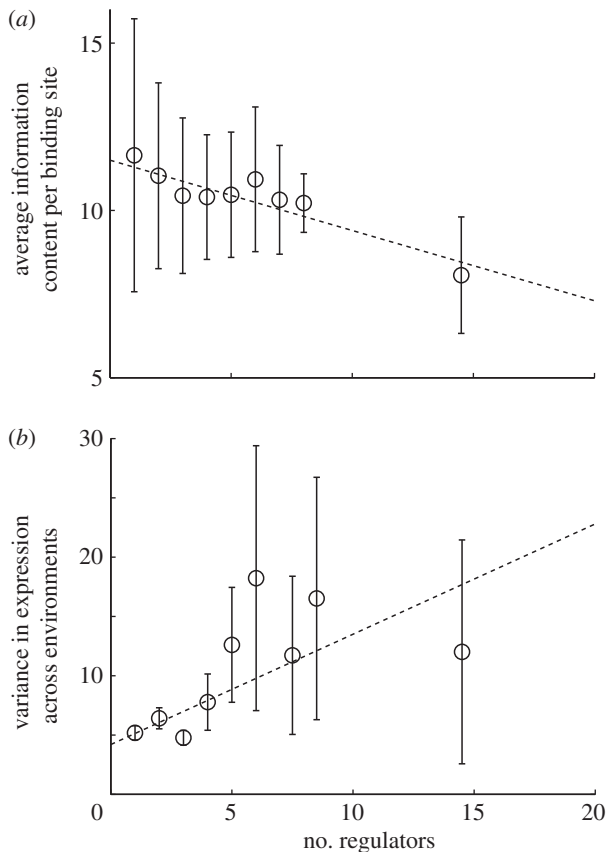
Our paper is structured as follows. We begin by inspecting empirical data on the relationship between motif specificity and regulatory complexity, in yeast. To help understand these data, we then construct a model of regulatory module evolution. Our analysis is based on a standard biophysical model for transcription factor binding, which we use to construct fitness landscapes for regulatory modules consisting of multiple binding sites that are selected to execute a given binding logic. We analyse the evolution of these modules in the limit of weak mutation (which is appropriate when considering binding sites at a particular target gene [14–16]), and we determine whether, at equilibrium under stabilizing selection [17], the binding sites that belong to a module are likely to be functional or non-functional (i.e. whether they are likely to be bound by their respective transcription factors, or not). We determine how the information content of the factors maintained in a module varies with the module's size, for fixed selection strength and population size. We initially focus on modules consisting of only two transcription factors, since this case can be understood analytically. We then employ evolutionary simulations to explore larger modules. We also analyse the impact of variation in the co-expression patterns of transcription factor proteins on the equilibrium size of a module. Finally, we study the generality of our results by relaxing model assumptions. We conclude by comparison with empirical data on transcription factor motifs and modules, within and between species.

## 2. Empirical trends in binding site specificities

Transcriptional control of a gene is often mediated by a regulatory module—that is, a set of upstream binding sites

associated with distinct transcription factors that, in combination, regulate the expression level of the target gene. Larger modules can assume a greater number of distinct states and thus execute more complex binding logics (figure 1). We focus on two pieces of empirical data concerning the complexity of regulatory modules and the specificity of the binding motifs they contain (figure 2). Our data come from yeast [18–20]. First, we studied the relationship between the number of binding sites that regulate a given target gene (that is, the gene's module size), and the average information content of the binding motifs in the module. We found a strong negative correlation between module size and the average information content of motifs: larger modules employ less-specific binding motifs (figure 2). We also found a negative correlation between module size and the variance in the information content of its regulatory motifs across target genes. Similar correlations were reported previously by Bilu & Barkai [13].

Our second empirical observation concerns the relationship between regulatory complexity and the variation in a gene's expression across environments. We found that genes whose expression varies substantially across environmental conditions tend to have a greater number of regulators than genes with little expression variation across environments (figure 2—Bilu & Barkai [13] also report a similar correlation). This observation supports our interpretation that more regulators provide greater regulatory complexity. These empirical results continue to hold when genes are separated according to whether they contain a TATA box [21,22]—or not (see the electronic supplementary material and figure S1). Furthermore, we repeated our analysis for the *Escherichia coli* transcription network [23] and found, once again, a strong negative correlation between the number of regulators of a gene and the average information content of its regulating binding motifs (see the electronic supplementary material, figure S2).



**Figure 2.** (a) The empirical relationship between the number of regulators of a gene and the information content of its regulatory binding sites, in the yeast transcription network [18–20]. Both the average information content of regulatory motifs (points) and its standard deviation (bars) decrease with the number of regulators. The dashed line is a linear fit to the data and shows a significantly negative slope ( $p < 2 \times 10^{-8}$ ). Points show the average information content of binding motifs, for target genes binned according to the number of regulators. (b) The empirical relationship between the number of regulators of a gene and the variance in the gene's expression across environments in the yeast transcription network [13, 18–20]. The variance in gene expression increases with the number of regulators. The dashed line is a linear fit to the data and shows a significantly positive slope ( $p < 8 \times 10^{-15}$ ). Points show the average variance in target gene expression across environments for target genes binned according to the number of regulators. Bin sizes were chosen so that each bin contains at least 10 target genes. Bars show the standard deviation in the information content of binding sites, taken across all genes with a given number of regulators. Bars extend 1 s.d. either side of the mean.

In summary, the empirical data in yeast reveal that (i) the more transcription factors that co-regulate a gene, the lower the information content of each factor's binding motif and (ii) the greater the expression variation of a gene across environments, the greater the number of transcription factors that co-regulate it. The second observation makes intuitive sense: genes that vary their expression substantially across environmental conditions generally require more complex regulation, and thus larger module size, in order to integrate and respond to different environmental signals. The first observation—that larger regulatory modules comprises individually less-specific regulatory motifs—is harder to understand. Why should genes controlled by complex regulatory modules use transcription factors with low information content? We seek to answer this question by considering the population genetics and evolution of regulatory modules.

### 3. Results

#### (a) Biophysical model of transcription factor binding

We use an established biophysical model of transcription-factor binding [24–28], which treats a binding site as a sequence of  $n$  consecutive nucleotides for which there is an associated consensus sequence (or set of sequences) that results in a minimum binding energy. Any set of  $n$  consecutive bases can be characterized by its number of 'mismatches', i.e. the number of nucleotide positions at which it differs from the consensus sequence. In the standard model, each such mismatch increases the binding energy of the sequence, compared with the consensus sequence, by an amount  $\epsilon$ . The increase in energy per-mismatch has been empirically measured to fall within the range 1 and 3  $k_B T$  [11,29]. The probability  $\pi_i$  that any given  $n$ -nucleotide sequence is bound by a transcription factor is determined by the number of mismatches,  $i$ , the binding energy per-mismatch,  $\epsilon$ , and the number of free transcription factor proteins in the cell,  $P$ , according to the following equation:

$$\pi_i = \frac{P}{P + \exp[\epsilon i]}. \quad (3.1)$$

We describe the consensus sequence of a binding motif by assuming that each of the  $n$ -nucleotide positions can be treated as having a degeneracy,  $r$ , which quantifies the (average) number of different bases that can appear at each position and still result in minimum binding energy. Thus, if  $r = 1$ , minimum binding energy is achieved only if each of the  $n$  nucleotides adopts a single specific base. If  $r = 2$ , minimum binding energy can be achieved if each of the  $n$  nucleotides adopts one of two bases, and so on. The average degeneracy for a given transcription factor can be calculated from the position-specific weight matrix (PSWM) of its binding site [12].

Increasing the average degeneracy  $r$  of a consensus sequence lowers the specificity of the motif, since a greater number of different nucleotide sequences result in minimum binding energy. Similarly, reducing the length of the consensus sequence,  $n$ , also decreases its specificity, since fewer nucleotides need to be matched to a specific base to minimize binding energy. In order to compare the specificities of different binding motifs with different lengths and degeneracies, we follow the approach used in earlier studies and measure the information content,  $I$ , of a PSWM [30,31],  $I = n \log_2[4/r]$ .

#### (b) Mutation and selection in the weak mutation limit

We use the probability of binding,  $\pi_i$  to construct the fitness landscape of a regulatory module. As in previous studies [28], we assume that fitness is a linear function of the probability that a binding site under selection is in fact bound. Thus, for a single binding site with  $i$  mismatched nucleotides, the fitness  $w_i$  is given by  $w_i = 1 - s(1 - \pi_i)$ . (We later generalize this to modules composed of multiple binding sites.) In the case of a single binding site, when  $\pi_i = 1$  the site is always bound producing fitness  $w_i = 1$ . The parameter  $s$  quantifies the reduction in fitness that occurs when the binding site is unbound, so that if  $\pi_i = 0$ , and the site is always unbound, we assign fitness  $w_i = 1 - s$ .

Following the approach used previously, we analyse binding site evolution at a single target gene in the limit of weak mutation [14–16]. This regime is realistic because the per-nucleotide mutation rate in both prokaryotes and

eukaryotes is low,  $\mu \sim O(10^{-8})$ , binding sites are typically short,  $n \sim O(10)$  and selection on conserved binding sites is sufficiently strong, with  $N_s \sim O(10)$ , where  $N$  is the population size. We later relax the assumption of weak mutation.

In the weak-mutation limit, evolution occurs through a series of selective sweeps, with new mutations arising only after earlier mutations have either fixed or gone extinct. We can calculate the equilibrium distribution  $F_i$  of binding sites with  $i$  mismatched nucleotides (see the electronic supplementary material). When there is no selection,  $s = 0$ , then  $F_i$  is just a binomial distribution whose mean is determined by the rates of mutations that increase or decrease  $i$ . However, when selection is present, i.e.  $s > 0$ , the equilibrium distribution  $F_i$  is bimodal, with one peak occurring at values of  $i$  for which  $\pi_i \sim 1$  and a second peak occurring at the neutral equilibrium,  $i = n(1 - r/4)$  (see electronic supplementary material, figure S3a).

### (c) A definition of functional binding

The bimodal form for the equilibrium distribution of mismatches provides a natural way to separate binding sites into ‘functional’ and ‘non-functional’ categories. This classification simplifies the analysis of modules containing multiple binding sites. We adopt a simple, operational definition of a functional binding site: all binding sites for which the probability of binding exceeds one half, i.e.  $\pi_i > 0.5$ , are defined as functional; all others are defined as non-functional. This is a natural definition because the probability of binding,  $\pi_i$ , is a sigmoidal function of the number of mismatches,  $i$ , with a fairly sharp threshold occurring at the value of  $i$  for which  $\pi_i = 0.5$  (the threshold value is given by  $i = \log[P]/\epsilon$ ).

In practice, this definition means that binding sites belonging to the mutation-selection peak of the bimodal distribution (figure 2a) are considered functional, whereas binding sites belonging to the drift peak of the bimodal distribution are considered non-functional. To ensure that a neutrally evolving region of DNA is not mislabelled as a functional binding site, we impose the additional constraint  $n(1 - r/4) > \log[P]/\epsilon$ . This constraint ensures that selection is able to differentiate between the functional (selective) peak and the non-functional (drift) peak in figure 2a. In essence, this condition excludes binding motifs with very low information content from our analysis. Furthermore, to account for the effects of spurious binding to low-information content binding sites, we analysed both the case in which the number of free transcription factor proteins,  $P$ , is independent of binding-site information, and the case when  $P$  decreases as information content decreases (owing to spurious binding), see later.

Given our definition of functional binding, the equilibrium probability,  $f$ , that a binding site will be functional is given by

$$f = \sum_{\{i | \pi_i > 0.5\}} F_i. \quad (3.2)$$

Electronic supplementary material, figure S3c shows how the probability a binding site is functional,  $f$ , depends upon the scaled strength of selection,  $N_s$ . This relationship displays a sharp, threshold behaviour [27]. Thus, as selection strength or population size increases, binding sites rapidly switch from having many mismatched nucleotides to having few mismatched nucleotides, so that the chance of binding rapidly switches from below one-half to above one-half.

### (d) Evolution of a single binding site

Before considering complex regulatory modules, we first analyse the evolution of a regulatory module composed of a single binding site. We consider a single binding site with information content  $I$  evolving in a population of size  $N$  under stabilizing selection of strength  $s$  for binding. As described earlier, we can determine the probability,  $f$  that it is functional in equilibrium. The information content of the binding site depends on its length  $n$  and average degeneracy  $r$ , which are independent parameters. However, any pair of values  $\{n, r\}$  that result in a given information content,  $I = n \log_2[4/r]$ , result in the same (or very close to the same) probability  $f$  that the site is functional (see the electronic supplementary material, figure S4). Thus, it is sufficient to know only the total information content of a binding motif in order to understand its evolution, as opposed to know both  $n$  and  $r$  independently. Therefore, we confine ourselves to discussing the information content of sites.

Typically, the strength of selection on binding sites is of the order  $N_s \sim 10$  [28,32]. Assuming  $N_s = 10$ , electronic supplementary material, figure S3b shows how the probability that the site is functional,  $f$ , depends on its information content,  $I$ . The figure also indicates the critical value of information that results in the binding site being functional with probability one-half; this occurs when  $I = 37.5$  bits. For values of information content more than 37.5 bits, the probability that the site is functional declines rapidly to zero; whereas a binding site with information less than 37.5 bits has the probability of being functional near 1. Therefore, a regulatory module consisting of a single binding site will probably be functional whenever  $I \leq 37.5$ , given selection of strength  $N_s = 10$ . This threshold should not be taken too literally, because  $N_s$  varies among binding sites and among species. However, for realistic parameter choices, this threshold information content is roughly comparable with the information content of empirical binding motifs [12]. The simple case of a single binding site, described in this section, forms a basis for comparison as we consider modules with multiple binding sites, below.

### (e) Evolution of regulatory modules composed of two binding sites

Next, we used our population-genetic model to study the evolution of regulatory modules composed of two binding sites, under selection for a specific binding logic. A pair of binding sites for two co-expressed transcription factors can be bound in four possible combinations. We assume that the regulated target gene produces fitness 1, if the two sites are bound in a ‘desired’ combination, and fitness  $1 - s$  otherwise. We refer to the pattern of bound combinations that are favoured by selection as the selected ‘binding logic’ of a module (figure 1).

For a module with two binding sites, three important binding logics may occur. Under selection for an AND binding logic, both factors must be simultaneously bound, so that the fitness of the regulated gene is 1 when both binding sites are bound and  $1 - s$  otherwise. Under selection for an OR binding logic, the fitness is 1 when either one or both of the binding sites is bound. Under selection for an XOR binding logic, the fitness is 1 when one but not both binding sites are bound. We study how the binding motifs in the



**Table 1.** Binding logics for modules with two binding sites.

logic gate	selected regulation	probability of selected binding
AND	$A$ and $B$ must both be bound for correct regulation	$\pi_i^A \pi_j^B$
OR	either $A$ or $B$ or both must be bound for correct regulation	$\pi_i^A + \pi_j^B - \pi_i^A \pi_j^B$
XOR	either $A$ or $B$ but not both must be bound for correct regulation	$\pi_i^A + \pi_j^B - 2\pi_i^A \pi_j^B$
no interaction	$A$ and $B$ contribute independently to correct regulation	$\pi_i^A + \pi_j^B$

regulatory region will evolve in response to the selected binding logic—including the possibility that some of the motifs might become non-functional (for example, if the selective binding logic is XOR then both transcription factors are present in the cell at all times).

We emphasize that the selected binding logic, shown schematically in figure 1, is distinct from the logic of physical interactions among factors bound in a regulatory region, which collectively determine whether the target gene is transcribed or not. Regardless of the interaction logic among bound factors, the selected binding logic in a module is what determines the long-term evolutionary outcome of the binding motifs [33].

For a pair of binding sites,  $A$  and  $B$ , with  $i$  and  $j$  mismatched nucleotides, respectively, the probability of desired binding for an AND regulatory module is given by  $\sigma_{ij}^{\text{AND}} = \pi_i^A \pi_j^B$ . For an OR regulatory module the probability of desired binding is  $\sigma_{ij}^{\text{OR}} = \pi_i^A + \pi_j^B - \pi_i^A \pi_j^B$ , and for an XOR module the probability of desired binding is  $\sigma_{ij}^{\text{XOR}} = \pi_i^A + \pi_j^B - 2\pi_i^A \pi_j^B$ . These three possibilities are summarized in table 1. Cooperativity between the two transcription factor proteins is neglected here, but considered below.

We have derived the equilibrium distribution of mismatched nucleotides ( $i, j$ ) for a two-site regulatory module, evolving under weak mutation (see the electronic supplementary material). Figure 3 shows contour plots for the probabilities,  $f_A$  and  $f_B$ , that each of the binding sites are functional in equilibrium, for different values of information content at each binding site. The figure also shows these contours for each of the three possible binding logics described above, with selective strength  $N_s = 10$  for the desired binding configurations.

Also shown for comparison in figure 3 is a contour map for two binding sites functioning independently in isolation of each other (figure 3a). Each such site corresponds to the case of a single binding site, as discussed in §3d.

Figure 3 illustrates the central result of our study: for all three possible binding logics, the range of information content for which both binding sites are functional (and thus retained) is much smaller than the information content of a functional, single-site module. Thus, our analysis predicts that functional binding sites (which are retained over evolution) in a two-site module, regardless of the selected binding logic, will tend to have less information than a binding site occurring in single-site module.

Although the analysis in figure 3 assumes that both transcription factors are constitutively co-expressed in the cell, our model can easily be generalized to the case in which pairs of factors are co-expressed only for some proportion of the time (see the electronic supplementary material). The results shown in figure 3 are qualitatively unchanged in

such a scenario (electronic supplementary material, figures S13–S15).

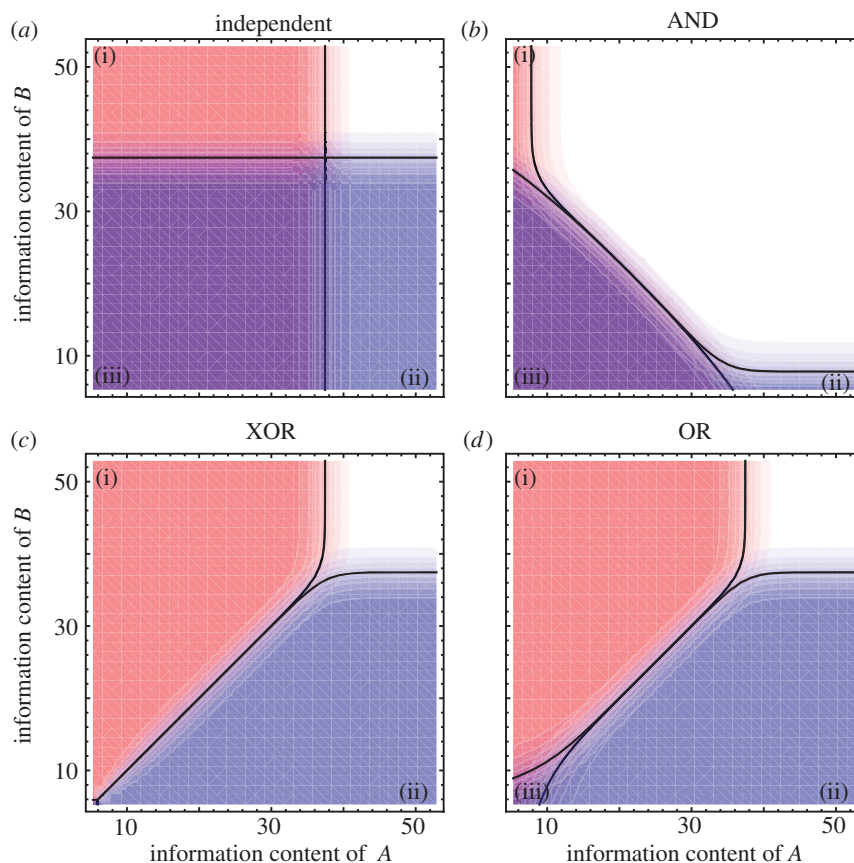
## (f) Evolution of regulatory modules composed of many binding sites

Our analysis of two-site regulatory modules can be extended to describe larger regulatory modules. We consider  $M$  co-expressed transcription factors, whose binding sites co-regulate a target gene. The group of  $M$  binding sites can be bound by their respective transcription factors in  $2^M$  possible combinations. Each combination of bound sites can, in turn, constitute a desirable or an undesirable pattern of gene regulation. As a result, there are  $2^{2^M}$  possible binding logics that can be executed by such a regulatory module. An example is illustrated in the electronic supplementary material, table S1. Analogously to the two-site case, we can construct a function  $\sigma_{i_1, i_2, \dots, i_M}$  to describe the probability that a set of transcription factor binding sites,  $\{A_1, A_2, \dots, A_M\}$ , with  $\{i_1, i_2, \dots, i_M\}$  mismatched nucleotides is bound in a desirable pattern, for a given regulatory binding logic. The equilibrium distribution for the number of mismatched nucleotides at each binding site can once again be found analytically. This expression is given in the electronic supplementary material, however the combinatorial explosion in the number of possible selected binding logics with module size means that an analytical exploration of modules with more than two binding sites quickly becomes impractical.

Instead, we performed simulations of evolution under weak mutation. We constructed regulatory modules in our simulations as follows: for each binding site, we drew a energy contribution per mismatch,  $\epsilon$ , and a number of proteins per cell,  $P$ , from a uniform distribution in the empirically determined ranges  $1 \leq \epsilon \leq 3$  and  $10^0 \leq P \leq 10^3$ , respectively. We also drew a binding site length,  $n$ , and average degeneracy,  $r$ , from a uniform distribution in the range  $5 \leq n \leq 40$  and  $1 \leq r \leq 4$ , respectively, with the additional constraint that we condition on  $n(1 - r/4) > \log[P]/\epsilon$ , to ensure that selection is able to differentiate between functional and non-functional sites.

We can extend our analysis of two-site modules by assuming a strict AND binding logic, or an OR binding logic, across all the binding sites in the module. Alternatively, we also explored random binding logics. To do this, we chose a binding logic uniformly from among the  $2^{2^M}$  possible logics.

Each module was then allowed to evolve until greater than  $10^2$  mutations fixed, to ensure that equilibrium had been reached, and the simulation was then stopped. Once stopped, a module was defined as functional if each of its composite binding sites is functional. We calculated the average information content of motifs, in all the functional



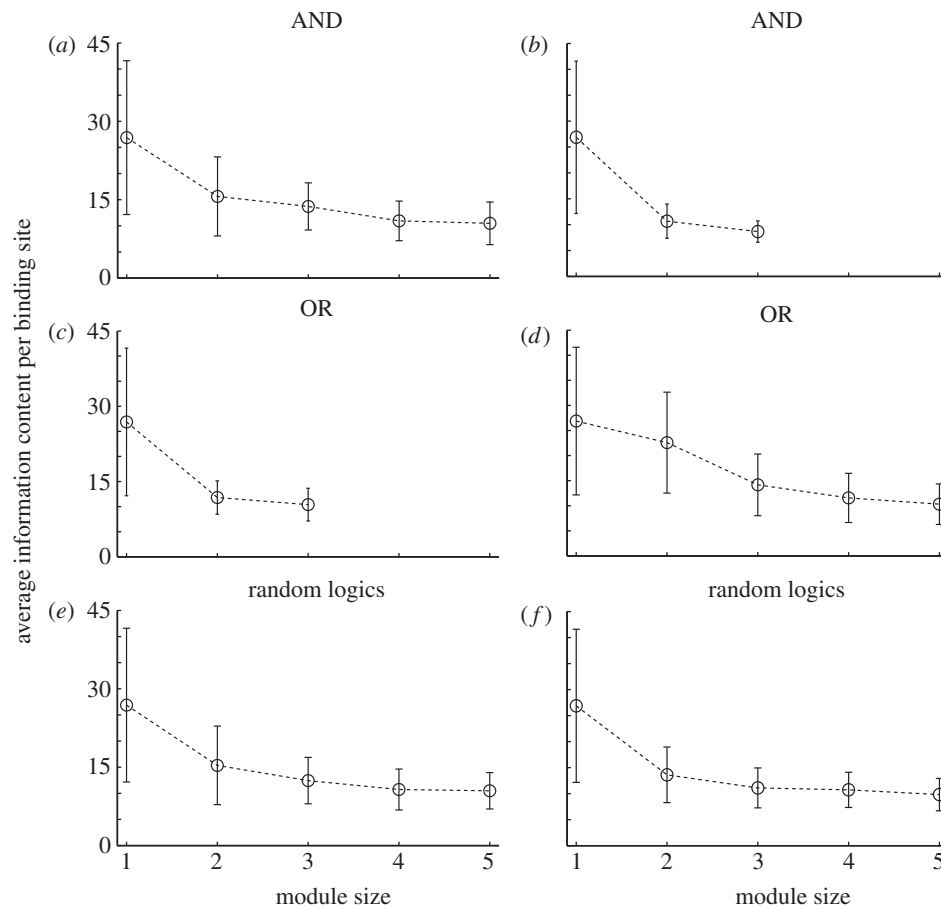
**Figure 3.** Regulatory modules containing two binding sites are composed of individually less specific binding sites, compared with a module composed of a single binding site. Each panel shows a pair of overlapping contour plots for the probability that binding site  $A$  is functional,  $f_A$ , and the probability that binding site  $B$  is functional,  $f_B$ . Solid black lines indicate the contours  $f_A = 0.5$  and  $f_B = 0.5$ . Region (i) indicates that only  $A$  is likely to be functional, region (ii) indicates that only  $B$  is likely to be functional and region (iii) indicates the overlap between regions (i) and (ii), meaning that both binding sites  $A$  and  $B$  are likely to be functional. White regions indicate that neither site is likely to be functional. All plots are generated with selection strength  $N_s = 10$ . The figures show: (a) probability of being functional for individual binding sites in isolation: region (iii), which occurs when both binding sites have  $I \lesssim 37.5$  bits, serves as a basis for comparison with two-site binding modules. (b) A two-site module with AND binding logic, so that both  $A$  and  $B$  are selected to be bound. Region (iii) is smaller than in the one-site case, indicating that functional binding sites maintained in the module will contain less information than in the one-site case. (c) A two-site module with XOR binding logic, so that  $A$  or  $B$  but not both are selected to be bound; again, the binding sites maintained in such a two-site module each have less information than in a single-site module. (d) A two-site module with OR binding logic, so that  $A$  or  $B$  is selected to be bound. Only a small region (iii) at low information content is visible. As a result, only the binding site with lower information content will typically remain functional. Information content is calculated by fixing degeneracy  $r = 1.6$ , and varying binding site length  $n$  to produce values of information content that coincide with the empirically observed range [12], with  $P = 100$  and  $\epsilon = 2$ . (Online version in colour.)

modules of a given size. Because the parameter space is large we constructed greater than  $10^8$  different regulatory modules, of each size. Of these modules, typically less than  $10^2$  are found functional in equilibrium.

Figure 4 shows the relationship between the average information content of binding sites and module size. We performed simulations using the randomly generated binding logics, described above, as well as modules executing strict AND and OR binding logics. In all the cases, we observe a decline in the average information content per binding site as module size increases. In addition, we observe a decline in the ensemble variance in information content of binding sites with module size. Thus, the region of parameter space in which functional modules are maintained by selection becomes smaller, and it includes lower information binding sites, as the module size (and hence regulatory complexity) increases. These results hold both for transcription factors that are constitutively co-expressed (figure 4a,c,e), and for transcription factors that are co-expressed only some proportion of the time (figure 4b,d,f).

### (g) Relaxation of model assumptions

Our analysis has relied on a number of assumptions regarding the mechanistic details of transcription factor binding to DNA. We now relax these assumptions to explore the generality of our results. Our analysis assumed that transcription factors bind their sites independently, whereas in reality transcription factors often bind cooperatively (or antagonistically). Such interactions can be incorporated into our model (see the electronic supplementary material). An example showing the effects of cooperativity on the evolution of two-site modules is given in the electronic supplementary material, figures S8–S10. We find that cooperative binding alters the quantitative range of motif information that can be maintained under stabilizing selection. However, cooperativity does not change our central conclusion: larger regulatory modules tend to be composed of individually less-informative sites. Likewise, when a transcription factor binds cooperatively at multiple copies of its binding site, we again find the same qualitative result (see the electronic supplementary material, figure S11). We also relaxed the



**Figure 4.** Information content of binding sites is predicted to decrease with module size, regardless of the selected binding logic. Points show the ensemble average information content per binding site in a module, and bars show the ensemble standard deviation (bar width is 2 s.d. either side of the mean). Panels top to bottom show modules with AND, OR and mixed (arbitrary) binding logics. (a,c,e) (100% overlap) corresponds to a model in which all all transcription factors are co-expressed, at all times. (b,d,f) (50% overlap) corresponds to a model in which any given pair of factors are co-expressed half the time. Monte Carlo simulations of binding site evolution in the weak selection limit were performed, as described in the main text. For each module size, replicate simulations were performed until at least  $10^2$  functional modules were produced. (Missing data points indicate that no functional modules were produced, after even  $10^6$  simulations.) All modules were evolved with selection strength  $N_s = 10$ . In all the cases, the average information content of the functional binding sites in a module, and the ensemble variance of information content among functional binding sites, decrease with module size,  $M$ .

assumption that the pool of transcription factors free to bind at a site,  $P$ , is independent of binding-site information, without altering our qualitative conclusions; our results are robust to this assumption as well (see the electronic supplementary material, figure S12).

Our analysis has assumed that stronger binding, in the selected binding logic, is always more beneficial. Nonetheless, several studies have found that binding sites with intermediate binding affinity are widespread [34–36], suggesting that selection might sometimes favour binding of intermediate affinity. Therefore, we reconsidered our analysis of two-site modules, in the case when selection favours binding of intermediate strength. We found that selection for intermediate-strength binding does not qualitatively alter our results on the relationship between module size and information content (see the electronic supplementary material, figure S7).

We expect regulatory evolution to occur in the regime of weak mutation, as has been commonly assumed in most biophysical studies [14–16]. Nonetheless, recombination can play an important role in *cis*-regulatory evolution [37]. Provided modules are not too large, and the weak mutation limit holds, we do not expect clonal interference or recombination to be important factors in determining the information

content of regulatory sites in our model. To test this expectation, we ran simulations with elevated per-base mutation rates ( $N\mu = 0.1$ ) including recombination. As shown in the electronic supplementary material, figure S16, we find the same relationship between module size and information content as found under weak mutation. Finally, we tested our results under different values of selection strength  $N_s$  (see the electronic supplementary material, figures S5–S6 and S17). As expected, stronger selection is able to maintain binding sites with greater information content, compared with weaker selection. But the basic trend of complex regulation by low-information binding sites holds for a range of selection strengths.

## 4. Discussion

We have shown that as the size of a regulatory module increases, the specificity of its constituent binding sites is expected to decrease, if they are to be maintained over evolution. This result is general in the sense that it does not depend on the binding logic under selection nor does it depend on the temporal expression pattern of the transcription factors participating in the module.

Our investigation was motivated by the empirical trend that greater regulatory complexity tends to be carried out by transcription factors with less-specific binding motifs, in yeast. Bilu and Barkai, who made similar empirical observations, suggested that this trend may be explained either by weaker selection on the expression of genes with highly variable expression, or by a tendency for multiple co-regulating binding sites to experience compensatory mutations. Our analysis does not rule out these explanations. However, we have shown that no assumptions about the strength of selection on gene expression, or about the epistatic effects of mutations among binding sites, are necessary to explain these empirical patterns. Rather, it is necessary only to consider the impact of stabilizing selection on the information content of binding sites maintained in a regulatory module.

Our results do not tell us how 'best' to construct a regulatory module for a given task, a question which has been addressed elsewhere [5,25,38]. Rather, our results hold regardless of the binding logic favoured by selection. In a certain light, our results make simple, intuitive sense: it is *all* of the nucleotide positions which belong to a regulatory module that form the target of selection, and so as module size increases, the maximum mutation rate at any given site, and hence its maximum information content, must decrease. What is not obvious, however, is that this effect should be so pronounced, given the information content and selection strengths typical of transcription factor binding sites. Yet, this is precisely what we find.

It is important to note that while we expect the trend of more complex regulation by lower specificity binding sites to hold quite generally, our results do not give an exact prescription for the information content of a given regulatory module, as this is influenced by the strength of selection on the target gene's expression, as well as the biophysics of transcription factor binding, which can vary among proteins and among species. For example, cooperativity between factors, selection for intermediate binding-site affinity, and multiple copies of the same binding site at a given target gene [34–36] all

influence of the details of transcription factor binding. In addition to these biophysical details, there are other evolutionary factors that can determine the specificity of binding motifs, such as binding site turnover, or selection against spurious binding elsewhere in the genome [12,31].

The results of our population-genetic analysis may also help to explain module composition across different species. In particular, prokaryotes are known to contain smaller transcriptional modules, on average, than eukaryotes (average module size 2.2 motifs in *E. coli* versus 3.0 motifs in *Saccharomyces cerevisiae* [18–20,23]). Moreover, lower eukaryotes tend to contain smaller modules than higher eukaryotes (average module size 6.6 in human [20,39]). According to our analysis, then, we would expect an inverse trend in the information content of motifs in these three species. Such a trend is indeed observed, with humans tending to use the least-specific motifs (11.2 bits per motif [20]), followed by yeast (12.1 bits per motifs [20]), and then by *E. coli* (14.9 bits per motif, [23]).

Higher eukaryotes must carefully orchestrate gene expression to produce the elaborate phenotypes associated with multi-cellularity. And even simple eukaryotes require complex regulation of genes to respond to different environmental conditions. In spite of this, many eukaryotic genes have noisy expression, and many of the transcription factors that regulate them bind weakly [40]. Our study suggests an evolutionary perspective on this phenomenon: complexity requires some sloppiness.

**Acknowledgements.** We thank Sridhar Hannenhalli, the Plotkin laboratory and two anonymous referees for constructive input, as well as C. Bergman for tweeting helpful references.

**Funding statement.** We acknowledge funding from the Burroughs Wellcome Fund, the David and Lucile Packard Foundation, the James S. McDonnell Foundation, the Alfred P. Sloan Foundation, grant no. D12AP00025 from the US Department of the Interior, grant no. W911NF-12-1-0552 from the US Army Research Office and grant no. RFP-12-16 from the Foundational Questions in Evolutionary Biology Fund.

## References

- Davidson EH *et al.* 2002 A genomic regulatory network for development. *Science* **295**, 1669–1678. (doi:10.1126/science.1069883)
- Eldar A, Elowitz MB. 2010 Functional roles for noise in genetic circuits. *Nature* **467**, 167–173. (doi:10.1038/nature09326)
- Guet CC, Elowitz MB, Hsing W, Leibler S. 2002 Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470. (doi:10.1126/science.1067407)
- Li H, Johnson AD. 2010 Evolution of transcription networks—lessons from yeasts. *Curr. Biol.* **20**, R746–R753. (doi:10.1016/j.cub.2010.06.056)
- Little JW. 2010 Evolution of complex gene regulatory circuits by addition of refinements. *Curr. Biol.* **20**, R724–R734. (doi:10.1016/j.cub.2010.06.028)
- Shen-Orr SS, Milo R, Mangan S, Alon U. 2002 Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68. (doi:10.1038/ng881)
- Tirosh I, Weinberger A, Carmi M, Barkai N. 2006 A genetic signature of interspecies variations in gene expression. *Nat. Genet.* **38**, 830–834. (doi:10.1038/ng1819)
- Tuch BB, Li H, Johnson AD. 2008 Evolution of eukaryotic transcription circuits. *Science* **319**, 1797–1799. (doi:10.1126/science.1152398)
- Lynch M. 2006 The origins of eukaryotic gene structure. *Mol. Biol. Evol.* **23**, 450–468. (doi:10.1093/molbev/msj050)
- Janga SC, Collado-Vides J, Babu MM. 2008 Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proc. Natl Acad. Sci. USA* **105**, 15 761–15 766. (doi:10.1073/pnas.0806317105)
- Lässig M. 2007 From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinform.* **8**(Suppl. 6), S7. (doi:10.1186/1471-2105-8-S6-S7)
- Stewart AJ, Plotkin JB. 2012 Why transcription factor binding sites are ten nucleotides long. *Genetics* **192**, 973–985. (doi:10.1534/genetics.112.143370)
- Bilu Y, Barkai N. 2005 The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biol.* **6**, R103. (doi:10.1186/gb-2005-6-12-r103)
- Berg J, Willmann S, Lässig M. 2004 Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.* **4**, 42. (doi:10.1186/1471-2148-4-42)
- Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. 2003 Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.* **3**, 19. (doi:10.1186/1471-2148-3-19)
- Sella G, Hirsh AE. 2005 The application of statistical physics to evolutionary biology. *Proc. Natl Acad. Sci. USA* **102**, 9541–9546. (doi:10.1073/pnas.0501865102)



17. Ludwig MZ, Bergman CM, Patel NH, Kreitman M. 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567. (doi:10.1038/35000615)
18. Lee TI *et al.* 2002 Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804. (doi:10.1126/science.1075090)
19. Harbison CT *et al.* 2004 Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99–104. (doi:10.1038/nature02800)
20. Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A. 2008 Jasp, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–D106. (doi:10.1093/nar/gkm955)
21. Basehoar AD, Zanton SJ, Pugh BF. 2004 Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**, 699–709. (doi:10.1016/S0092-8674(04)00205-3)
22. Rando OJ, Winston F. 2012 Chromatin and transcription in yeast. *Genetics* **190**, 351–387. (doi:10.1534/genetics.111.132266)
23. Gama-Castro S *et al.* 2011 Regulondb version 7.0: transcriptional regulation of *Escherichia coli* k-12 integrated within genetic sensory response units (sensor units). *Nucleic Acids Res.* **39**, D98–D105. (doi:10.1093/nar/gkq1110)
24. Bintu L, Buchler N, Garcia H, Gerland U, Hwa T, Kondev J, Phillips R. 2005 Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15**, 116–124. (doi:10.1016/j.gde.2005.02.007)
25. Buchler NE, Gerland U, Hwa T. 2003 On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA* **100**, 5136–5141. (doi:10.1073/pnas.0930314100)
26. Chu D, Zabet NR, Mitavskiy B. 2009 Models of transcription factor binding: sensitivity of activation functions to model assumptions. *J. Theor. Biol.* **257**, 419–429. (doi:10.1016/j.jtbi.2008.11.026)
27. Gerland U, Hwa T. 2002 On the selection and evolution of regulatory DNA motifs. *J. Mol. Evol.* **55**, 386–400. (doi:10.1007/s00239-002-2335-z)
28. Mustonen V, Kinney J, Callan Jr CG, Lässig M. 2008 Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl Acad. Sci. USA* **105**, 12 376–12 381. (doi:10.1073/pnas.0805909105)
29. Gerland U, Moroz JD, Hwa T. 2002 Physical constraints and functional characteristics of transcription factor–DNA interaction. *Proc. Natl Acad. Sci. USA* **99**, 12 015–12 020. (doi:10.1073/pnas.192693599)
30. D'haeseleer P. 2006 What are DNA sequence motifs? *Nat. Biotechnol.* **24**, 423–425. (doi:10.1038/nbt0406-423)
31. Sengupta AM, Djordjevic M, Shraiman BI. 2002 Specificity and robustness in transcription control networks. *Proc. Natl Acad. Sci. USA* **99**, 2072–2077. (doi:10.1073/pnas.022388499)
32. He BZ, Holloway AK, Maerkl SJ, Kreitman M. 2011 Does positive selection drive transcription factor binding site turnover? a test with *Drosophila* cis-regulatory modules. *PLoS Genet.* **7**, e1002053. (doi:10.1371/journal.pgen.1002053)
33. Tsong AE, Tuch BB, Li H, Johnson AD. 2006 Evolution of alternative transcriptional circuits with identical logic. *Nature* **443**, 415–420. (doi:10.1038/nature05099)
34. Essien K, Vigneau S, Apreleva S, Singh LN, Bartolomei MS, Hannehalli S. 2009 Ctf binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol.* **10**, R131. (doi:10.1186/gb-2009-10-11-r131)
35. He X, Duque TSPC, Sinha S. 2012 Evolutionary origins of transcription factor binding site clusters. *Mol. Biol. Evol.* **29**, 1059–1070. (doi:10.1093/molbev/msr277)
36. Tanay A. 2006 Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972. (doi:10.1101/gr.5113606)
37. Paixão A, Azevedo RBR. 2010 Redundancy and the evolution of cis-regulatory element multiplicity. *PLoS Comput. Biol.* **6**, e1000848. (doi:10.1371/journal.pcbi.1000848)
38. Lynch M. 2007 The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* **8**, 803–813. (doi:10.1038/nrg2192)
39. Bauer T, Eils R, König R. 2011 RIP: the regulatory interaction predictor—a machine learning-based approach for predicting target genes of transcription factors. *Bioinformatics* **27**, 2239–2247. (doi:10.1093/bioinformatics/btr366)
40. Blake WJ, Balázs G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, Cantor CR, Walt DR, Collins JJ. 2006 Phenotypic consequences of promoter-mediated transcriptional noise. *Mol. Cell* **24**, 853–865. (doi:10.1016/j.molcel.2006.11.003)