The influence of promoter architectures and regulatory motifs on gene expression in *Escherichia coli*

Mattias Rydenfelt [1,2], Hernan G. Garcia [3], Robert Sidney Cox III [4], and Rob Phillips[5,6*]
1 Department of Physics, California Institute of Technology, Pasadena, CA, United States of America,
2 Integrative Research Institute for the Life Sciences and Institute for Theoretical Biology, Humboldt University, Berlin, Germany.
3 Joseph-Henry Laboratories of Physics, Princeton University, Princeton, NJ United States of America.
4 Department of Chemical Science and Engineering, Kobe University, Kobe, Japan.
5 Department of Applied Physics, California Institute of Technology, Pasadena, CA, United States of America.
6 Division of Biology, California Institute of Technology, Pasadena, CA, United States of America.
∗ E-mail: phillips@pboc.caltech.edu

Abstract

The ability to regulate gene expression is of central importance for the adaptability of living organisms to changes in their external and internal environment. At the transcriptional level, binding of transcription factors (TFs) in the promoter region can modulate the transcription rate, hence making TFs central players in gene regulation. For some model organisms, information about the locations and identities of discovered TF binding sites have been collected in continually updated databases, such as RegulonDB for the well-studied case of *E. coli*. In order to reveal the general principles behind the binding-site arrangement and function of these regulatory architectures we propose a random promoter architecture model that preserves the overall abundance of binding sites to identify overrepresented binding site configurations. This model is analogous to the random network model used in the study of genetic network motifs, where regulatory motifs are identified through their overrepresentation with respect to a "randomly connected" genetic network. Using our model we identify TF pairs which coregulate operons in an overrepresented fashion, or individual TFs which act at multiple binding sites per promoter by, for example, cooperative binding, DNA looping, or through multiple binding domains. We furthermore explore the relationship between promoter architecture and gene expression, using three different genome-wide protein copy number censuses. Perhaps surprisingly, we find no systematic correlation between the number of activator and repressor binding sites regulating a gene and the level of gene expression. A position-weight-matrix model used to estimate the binding affinity of RNA polymerase (RNAP) to the promoters of activated and repressed genes suggests that this lack of correlation might in part be due to differences in basal transcription levels, with repressed genes having a higher basal activity level. This quantitative catalogue relating promoter architecture and function provides a first step towards genome-wide predictive models of regulatory function.

Introduction

One of the most impressive accomplishments in molecular biology over the past half-century has been the mapping of thousands of gene interactions to create genetic networks for a broad collection of organisms. Such maps have made it possible to qualitatively understand how groups of genes can together provide important functionality. Still, the genetic network descriptions leave us with a picture of the regulatory landscape that is not quantitatively predictive. Although impressive, genetic networks do not provide us with all the information necessary to make concrete predictions, such as the number of proteins produced of a given kind under particular environmental conditions, not the least because the notions of 'activation' and 'repression' are often inherently verbal rather than quantitative. The amount of activation or repression achieved by transcription factors (TFs) can vary by many orders of magnitude, depending

on how tightly TFs bind to the promoter of interest [1] and many other factors. Moreover the resulting response curves depend on *promoter architecture*, i.e. the particular configuration of TF binding sites. For example, a repressor that blocks a promoter through DNA looping (e.g. LacI) has been shown to have a steeper response curve than its unlooped counterpart [2]. Furthermore, genetic networks do not tell if the TF that is supposed to regulate a gene is actually present in the cell at all, which might not be the case if it is inactivated through nucleosomal organization or by chromatin remodeling complexes [3, 4].

For genetic networks to be predictive tools in biology they need to be be augmented with quantitative descriptions of the census of regulatory players. Our goal with the present paper is to take a step in this direction by studying the role of promoter architectures in transcriptional regulation, from a genome-wide point of view. No organism offers a better opportunity to do so than *E. coli*, which after more than half a century of intense study demonstrates the most well understood regulatory network. Through ambitious efforts many cold and hard facts about transcriptional regulation in *E. coli* have been collected and made easily accessible in databases like RegulonDB [5] and EcoCyc [6]. These contain information including, but not limited to, which TFs regulate different operons, where they bind to promoters, and their regulatory effect (activation or repression). All of these features play an important role in transcriptional regulation. A TF which binds cooperatively to multiple binding sites, either through direct contact or DNA looping, provides a steeper regulatory response, typically reported by Hill coefficients, than TFs binding just a single site [7]. The position of binding sites play an equally important role. In experiments where a single repressor binding site has been systematically moved along the promoter region [8–11], the repression shows a clear dependence on position, interestingly featuring a 10-11 bp modulation following the periodicity of the DNA helix.

In this paper we study both the positions and multiplicities of TF binding sites in *E. coli*, for the 2500 or so known TF-DNA interactions in RegulonDB 8.5. A challenge inherent in using RegulonDB, EcoCyc, or any other biological database as primary information source is that the data is inevitably incomplete. More than half of the genes in *E. coli* still lack any regulatory annotation, including important genes such as those responsible for mechanosensation. We must therefore be cautious when interpreting our results. Whereas there is no obvious reason that, for example, binding site positions are biased, the absolute number of binding sites is almost certainly underestimated. This assertion is supported by the fact that the rate of newly discovered TF binding sites does not show any sign of slowing down, thanks to the advent of powerful techniques such as ChIP-seq [12] and Sort-Seq [13]. A healthy skepticism from the reader is thereby encouraged and the results should be viewed as provisional until more of the underlying regulatory facts are in hand.

We view the work presented here as a step towards using promoter architectures to give a more detailed understanding of transcriptional regulation than can be given by a genetic network map alone. Hopefully these findings can also provide valuable input for the theoretical dissection of transcription regulation, which has shown increasing capability to make distinct predictions for the response function of different promoter architectures [14–17]. Perhaps most importantly, the analysis presented here shows how far short the current factual understanding of regulatory architectures and measured expression levels falls from serving as a predictive framework, and thus should be seen as a call for higher predictive expectations and a more rigorous treatment of the relation between regulatory architecture and input-output functions.

Models

Random promoter architecture model

Following the classic method of random graphs [18], biological networks have been compared to randomly constructed networks to find *network motifs*, corresponding to recurring patterns in the connections between genes, which are overrepresented compared to a random graph [19]. One well-studied network motif is the feed-forward loop, where a single gene is regulated by two TFs, and in addition one of

the TFs regulates the other. Network motifs are presumedly selected for in biological systems because of functionality they provide, for example, robustness against concentration fluctuations of regulatory molecules. We similarly use a random assignment of TFs to create a null model of promoter architecture, and identify overrepresented promoter architectures motifs deviating from this expectation. For this we need to introduce a *random promoter architecture model*, to be used as reference for identifying overrepresented reported promoter architectures.

TF binding sites can be both lost and gained due to the steady pace of mutations across the genome. If we assume that these mutations occur randomly and uniformly across the genome, then in the absence of selection any specific distribution of a given number of TF binding sites over a set of operons would be as probable as any other. If a certain class of promoter architecture occurs more frequently in real regulatory networks than in this null model, we expect them to encode biological functions which are advantageous. The simple approach we will adopt to implement a random promoter architecture is therefore to imagine all binding sites reported in RegulonDB as being "sprinkled" over all operons with *uniform* probability. The mathematical implications of this simple postulate will be developed here, saving for the Results section the task of identifying promoter architecture motifs and how they differ from this simple null model.

As a first application of the random promoter architecture model we will look at the distribution of number of binding sites per operon. Throughout this work we will consider binding sites for a given TF as indistinguishable even if they have a different DNA sequence. Additionally, we define an operon as a cluster of transcriptional units, containing one or more protein coding sequences and one or more promoters which initiate transcription in the same direction to create mRNAs which carry the protein coding sequences. While some transcriptional units express RNA with no protein coding sequences, we ignore these cases for the time being. For this model we also explicitly neglect cases where one binding site can regulate more than one operon. In fact, only a small number of binding sites are known to regulate multiple operons (9% in RegulonDB 8.5). With these assumptions we can describe the random model of promoter binding site architecture.

There are 2871 TF binding sites listed in RegulonDB 8.5, and $N_{op} = 2642$ operons. We first consider the distribution of a single type of TF binding site with N_{bs} copies. The probability $P_{bs}(m; N_{bs})$ of a given operon to have exactly m binding sites assigned is described by the binomial distribution

$$P_{bs}(m; N_{bs}) = \binom{N_{bs}}{m} \left(\frac{1}{N_{op}}\right)^m \left(1 - \frac{1}{N_{op}}\right)^{N_{bs}-m} \qquad (m = 0, 1, 2, \ldots) \tag{1}$$

Here $\binom{N_{bs}}{m}$ is the number of ways to choose a set of m binding sites from the pool of N_{bs} sites, $(1/N_{op})^m$ is the probability that they are all assigned to a given operon, and $(1 - 1/N_{op})^{N_{bs}-m}$ is the probability that the rest of the bindings sites are assigned to the other operons.

Since the probability is small for a binding site to be assigned a particular operon, namely $1/N_{op}$, the binomial distribution can be approximated by a Poisson distribution with mean $\lambda = \frac{N_{bs}}{N_{op}}$. For the numbers of N_{bs} and N_{op} given by RegulonDB 8.5 the Poisson approximation is valid to within 1 % for $m \lesssim 10$, which covers 98 % of the operons in the dataset. However, for very highly regulated operons the Poisson approximation should not be used.

We can generalize the distribution in to incorporate several types of binding sites [see Eq. (2)], say activators or repressors, or different particular TFs. Since all binding sites are assumed to be independently distributed, the probability distribution $P_{2\,bs}(m_1, m_2; N_{bs}^{(1)}, N_{bs}^{(2)})$ for an operon to end up with m_1 binding sites (from a total of $N_{bs}^{(1)}$) of one type and m_2 binding sites (from a total of $N_{bs}^{(2)}$) of a second type will

be given by an independent product of two binomial distributions

$$P_{2\,bs}(m_1, m_2; N_{bs}^{(1)}, N_{bs}^{(2)}) = P_{bs}(m_1; N_{bs}^{(1)}) P_{bs}(m_2; N_{bs}^{(2)}) \tag{2}$$

$$= \binom{N_{bs}^{(1)}}{m_1} \binom{N_{bs}^{(2)}}{m_2} \left(\frac{1}{N_{op}}\right)^{m_1+m_2} \left(1 - \frac{1}{N_{op}}\right)^{N_{bs}^{(1)}+N_{bs}^{(2)}-m_1-m_2}. \tag{3}$$

Several TFs in *E. coli* preferentially bind to multiple binding sites at a given promoter, for example NarP binds to two sites or more at 10 out of 11 regulated operons according to RegulonDB 8.5. Such examples of operator multiplicity can occur due to cooperative contacts between the protein copies, or due to other reasons such as multiple transcription start sites and other TF interactions. How can we rigorously define the level at which binding sites of a TF cooccur around a promoter? Simply looking at the absolute number of operons with multiple binding sites is not a good measure of clustering of a TF at a promoter, as it does not take into account the total number of binding sites available. A *global TF* [20, 21] with hundreds of binding sites will likely bind at multiple sites at several promoters simply by chance.

We use the random promoter architecture model to derive the probability $P_{co}(M; N_{bs})$ for M operons to be regulated by at least two binding sites for the same TF, as a function of the total number of binding sites N_{bs} for that TF ($N_{bs} \geq 2M$). To find the number of ways N_{bs} binding sites can be distributed over N_{op} operons, with two or more sites at M of these, we first choose M operons, in any of $\binom{N_{op}}{M}$ ways, and assign two binding sites to each of them. See illustration Fig. 1(A) for a schematic description of the model. Next we put k of the remaining $N_{bs} - 2M$ binding sites into k of the remaining $N_{op} - M$ operons (i.e. one binding site per operon), which we can choose in $\binom{N_{op}-M}{k}$ ways. Finally we put the remaining $N_{bs} - 2M - k$ binding sites into the M operons, which already have two binding sites, however we want. The number of ways this can be done equals the number of nonnegative integer solutions to the equation $x_1 + x_2 + \cdots + x_M = N_{bs} - 2M - k$, a famous problem from combinatorics which is equivalent to the number of ways of placing $(N_{bs} - 2M - k)$ identical balls into M bins, with $\binom{N_{bs}-M-k-1}{M-1}$ solutions. To find the probability $P_{co}(M; N_{bs})$ we sum over k and divide by the total number of ways to distribute N_{bs} binding sites over N_{op} operons, which according to the same argument as above is given by $\binom{N_{bs}+N_{op}-1}{N_{bs}}$, resulting in (for $M \leq \min(N_{op}, \lfloor N_{bs}/2 \rfloor)$)

$$P_{co}(M; N_{bs}) = \frac{1}{\binom{N_{bs}+N_{op}-1}{N_{bs}}} \sum_{0 \leq k \leq \min(N_{op}-M, N_{bs}-2M)} \tag{4}$$

$$\times \binom{N_{op}}{M} \binom{N_{op}-M}{k} \binom{N_{bs}-M-k-1}{M-1}. \tag{5}$$

By using the continuous definition of binomial coefficient $\binom{x}{y} = \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}$ where $\Gamma(x)$ is the gamma function [22], Eq. (5) gives us the right probability also for $M = 0$, namely $P_{co}(0; N_{bs}) = \binom{N_{op}}{N_{bs}} / \binom{N_{bs}+N_{op}-1}{N_{bs}}$ when $N_{op} \geq N_{bs}$, and $P_{co}(0; N_{bs}) = 0$ when $N_{op} < N_{bs}$.

We can generalize this problem and analyze whether binding sites of a pair of TFs tend to cluster together. TF pairs which coregulate operons more frequently than suggested by the random promoter architecture model are more likely to have related biological function. Let $N_{bs}^{(1)}$ and $N_{bs}^{(2)}$ be the number of binding sites for two different types of TFs. As above we start by choosing M operons where we put one binding site of each kind. Next we put the remaining $N_{bs}^{(1)} - M$ binding sites of first type into the M "shared" operons plus an additional of k operons, which we can choose in $\binom{N_{op}-M}{k}$ ways, with at least one binding site in each. Fig. 1(B) gives a schematic of this procedure. The number of ways this can be done equals the number of integer solutions to the equation below with the given constraints

$$\underbrace{x_1 + \ldots + x_M}_{x_i \geq 0} + \underbrace{x_{M+1} + \ldots + x_{M+k}}_{x_i \geq 1} = N_{bs}^{(1)} - M. \tag{6}$$

After subtracting k from both sides one realizes that the number of solutions to this equation equals the number of nonnegative integer solutions to the simpler Eq. (7), with $\tilde{x}_i = x_i - 1$,

$$\underbrace{x_1 + \ldots + x_M + \tilde{x}_{M+1} + \ldots + \tilde{x}_{M+k}}_{x_i \geq 0} = N_{bs}^{(1)} - M - k, \tag{7}$$

which is given by $\binom{N_{bs}^{(1)}-1}{M+k-1}$. Next we distribute the remaining $N_{bs}^{(2)} - M$ binding sites of the second kind onto any of the operons, except from the k operons dedicated for binding sites of the first kind only. This can be done in $\binom{N_{op}-k+N_{bs}^{(2)}-M-1}{N_{bs}^{(2)}-M}$ ways. We find the probability $P_{2\,co}(M; N_{bs}^{(1)}, N_{bs}^{(2)})$ for M operons ($M \leq \min(N_{op}, N_{bs}^{(1)}, N_{bs}^{(2)})$) to be regulated by at least one binding site of each type by summing over k and dividing by the total number of binding site arrangements, which is given by $\binom{N_{op}+N_{bs}^{(1)}-1}{N_{bs}^{(1)}}\binom{N_{op}+N_{bs}^{(2)}-1}{N_{bs}^{(2)}}$, resulting in

$$\begin{aligned} P_{2\,co}(M; N_{bs}^{(1)}, N_{bs}^{(2)}) = & \frac{1}{\binom{N_{op}+N_{bs}^{(1)}-1}{N_{bs}^{(1)}}\binom{N_{op}+N_{bs}^{(2)}-1}{N_{bs}^{(2)}}} \times \\ & \sum_{k=0}^{\min(N_{bs}^{(1)}-M, N_{op}-M)} \binom{N_{op}}{M}\binom{N_{op}-M}{k}\binom{N_{bs}^{(1)}-1}{M+k-1} \\ & \times \binom{N_{op}-k+N_{bs}^{(2)}-M-1}{N_{bs}^{(2)}-M}. \end{aligned} \tag{8}$$

As a sanity check we use MATHEMATICA to see that the probabilities add up to one.

We can also compare with earlier work, which solved essentially the same problem but under the assumption that binding sites, even of the same kind, are distinguishable [23]. However, as long as the probability is small that two binding sites regulate the same operon the two methods will give similar results, just like Fermi-Dirac statistics approaches Boltzmann statistics in dilute systems [24].

A different method to identify TF cooperativity based on mutual information from ChIP data was used in [25]. The advantage of the random promoter architecture model is that it resolves biasing due to differences in number of TF binding sites, and allows us to determine both the expected number of coregulated operons and also the associated p-value of any given observation: i.e. the probability of an equal or more extreme outcome with respect to the random promoter architecture model. This will become useful in the Results section where we want to identify TF binding motifs in the reported distributions from RegulonDB.

Linear energy model of RNAP-DNA binding

The binding affinity of RNAP to the promoter of a gene is determined by the nucleotide sequence of the promoter and has a strong influence on the transcription rate of the gene [17]. The more effectively a promoter can recruit RNAP and initiate open complex formation, the higher the transcription rate of the gene will be. Creating a predictive map between DNA sequence and RNAP binding affinity is a problem which has received much attention [13, 26–29]. One of the simplest but yet most successful approaches to the modeling of RNAP-DNA (or TF-DNA) interactions is to assume *independent* energy contributions from each individual nucleotide in the binding sequence. Under this linear assumption the total binding energy of a sequence S can be expressed as a simple matrix trace

$$E(S) = \sum_{i=1}^{L} \sum_{j=A,C,T,G} M_{i,j} S_{j,i} = \mathrm{Tr}(MS), \tag{9}$$

where $S_{A/C/T/G,i} = 1$ if the identity of the base at nucleotide position i in the sequence is given by $A/C/T/G$ and otherwise $S_{A/C/T/G,i} = 0$, $M_{i,A/C/T/G}$ represents the energy contribution at position i for base $A/C/G/T$ respectively, and L is the length (in base pairs) of the binding sequence.

Despite having considered all promoter types up until this point, we now focus our attention on promoters associated with the sigma factor σ^{70}, which is called the "housekeeping" sigma factor of *E. coli*. The RNAPσ^{70} complex has two binding domains which may interact with the promoter at the -35 and -10 signals upstream of the transcription start site (+1) [30]. In this study we compute the binding energy from these two signals separately using the energy matrix $M_{i,j}$ of Brewster et al. [26]. Starting with two position weight matrices for the two signals, we place the "-10 box" signal at the -9, -10, or -11 position and the "-35 box" signal 22-24 bp upstream from the -10 box. Thus we allow 9 possible configurations for each pair of -10 and -35 boxes, and choose the one with the lowest binding energy according to the reference energy matrix [26].

The commonly used *occupancy hypothesis* [9,14], states a linear relationship between the transcription rate of a gene and the probability of its promoter being occupied by RNAP. This probability is, according to the Boltzmann distribution, proportional to $e^{-E(S)/k_BT}$ for systems in (quasi)equilibrium, an approximation which can be made if RNAP homogenizes throughout the cell at a much higher rate than that at which they are being produced. Despite its simplicity in ignoring details of open complex formation and promoter escape rate, the occupancy hypothesis has proved surprisingly successful in many different settings [1,9,26].

Results

How many genes do TFs regulate?

Genes of related biological function are often coregulated. For example a flagellum in *E. coli* consists of roughly forty different proteins [31] present at precise copy number ratios. Not all the flagellar genes are contained within the same operon, but instead these forty coding regions are transcribed from roughly ten different operons [6] (Fig. 2(A)). Coregulation of these operons allows the flagellar proteins to be expressed at precise ratios, a task which is handled by the TFs FlhC and FlhD in *E. coli*. However, other biological functions correlate with the production of flagella. For example production of sugar receptors in the cell membrane such as MglBAC, necessary for chemotaxis, are also regulated by FlhC and FlhD.

In general we expect "correlated genes" to be regulated by the same TFs, and the question we will address in this section is: How many correlated genes are typically regulated by the same TFs? In Fig. 2(B) we show the number of operons that are regulated by the same TF as reported by RegulonDB 8.5. By counting the number of coding sequences within each operon we also display the number of regulated genes per TF. Finally, in Fig. 2(B) we show the number of binding sites for each TF. The numbers provide a lower estimate of the actual *E. coli* regulatory network, acknowledging the fact that not all binding sites have yet been discovered. The figure reveals two almost separate groups of TFs: a large number of *specific TFs* which regulate only a few operons, and a mere handful of *global TFs* [20,21] regulating up to a hundred operons [see Table 1] each. Half of all TFs regulate two operons or less, suggesting that, unlike the construction of flagella, many operons in *E. coli* are not strongly correlated and encode all of the proteins necessary for a particular phenotype. For example, in response to varying levels of copper in the cytoplasm ComR reportedly regulates only one single gene, *bhsA*, which alters the outer cell membrane permeability for copper [32]. Global TFs, which regulate core activities in the cell, for example metabolic pathways (e.g. CRP) or the rRNA of the translational machinery (e.g. Fis), are the exceptions to this rule. Despite the small number of global TFs, these are involved in roughly half of all reported regulatory interactions.

To evaluate the regulatory complexity of a promoter, we can conversely consider the number of TFs regulating each operon. The more TFs regulating an operon, the more specific its response might be

to various cellular conditions. Note that here we will consider operons regulated by any of the *E. coli* σ subunits. As a result, the numbers below are certain to change if restricting the analysis to σ^{70}. In Fig. 3 we show the number of TF interactions and number of different TF types regulating operons as reported by RegulonDB 8.5. The average number of TF binding sites per operon is only 1.1, but climbs to 3.5 when excluding operons without known regulatory interactions. This observation suggests that data in RegulonDB is, to some extent, collected "one operon at a time" where the attention of the research community is focused on one operon before moving to the next one. There is an approximately exponential decrease (see fit) in the reported number of operons as a function of the number of their regulatory interactions. To see if the binding site multiplicity profiles differ between global TFs and specific TFs we show in Supporting Information Fig. S2 the profiles for these two groups separately, but find no significant differences. It is perhaps surprising that even for such a well studied organism as *E. coli* more than half of the genes still lack any regulatory annotation. Among these unannotated genes we find important examples such as the genes responsible for mechanosensation *mscS, mscL, mscK, ynal, ybio* and *ybdG*. Preliminary results from our lab based on the method of Sort-seq [13] show that at least some of these genes might in fact be regulated. Other notable genes lacking regulatory annotation include: *lpp*, a lipoprotein believed to be one of the most abundant proteins in *E. coli* [33]; *rep*, a helicase required for genomic replication [34] ; *kdpD* and *nhaB*, genes related to regulation of potassium [35] and sodium [36] levels in the cell. Nevertheless, it is still clear that many genes in *E. coli* do not strictly depend on TFs to be transcribed. This is in contrast with eukaryotic transcription where general TFs are necessary for the promoter recognition and transcription initiation process.

We can compare the observed distribution of number of TF interactions per operon with the random promoter architecture model [see Models]. Looking at Fig. 4(A) we see some notable differences between the random promoter architecture model and the observed distribution. A larger number of operons are reported as unregulated in RegulonDB 8.5 than expected from the random promoter architecture model. Some TFs tend to bind to multiple sites per operon, which could result in a higher number of unregulated operons as compared to the random architecture model. We will address this in more detail below. Another explanation for the high number of unregulated operons could simply be that RegulonDB 8.5 is inherently biased and reports a higher fraction of unregulated operons than the actual value. The logic behind this hypothesis is that those operons for which there are known binding sites correspond in general to those that have been studied carefully, whereas many operons with no annotated binding sites simply have not been studied in detail. To consider the later possibility we modified the random promoter architecture model to exclude operons with no known regulatory interactions. In this case we update the prediction of the random promoter architecture model [Eq (2)] by first assigning one binding site to each of $N_{op}^{(reg)}$ regulated operons. Then we randomly distribute the remaining $N_{bs} - N_{op}^{(reg)}$ binding sites on the $N_{op}^{(reg)}$ operons, as in Eq (2), leading to

$$P_{bs}(m; N_{bs}) = \binom{N_{bs} - N_{op}^{(reg)}}{m-1} \left(\frac{1}{N_{op}^{(reg)}}\right)^{m-1} \left(1 - \frac{1}{N_{op}^{(reg)}}\right)^{N_{bs} - N_{op}^{(reg)} - m + 1} \quad (m = 1, 2, 3, \ldots). \tag{10}$$

In Fig. 4(A) we now observe an overrepresentation of operons regulated from a single binding site, compared to the random promoter architecture model (compare black and blue dashed lines). This supports the idea that *E. coli* generally favors simple regulatory strategies when possible. In Supporting Information Fig. S3 we show that this conclusion holds separately for both global TF and specific TF binding sites. There is also a small group of highly regulated operons. For example *gadAXW*, coding for genes in the acid resistance system [37], is regulated by 35 TF binding sites. The operon *csgDEFG*, coding for genes that regulate the assembly and transport of extracellular amyloid fibres (known as Curli) [38], is regulated by 33 TF binding sites. Finally the operon *glpTQ*, shown schematically in Fig. 4(B), coding for genes responsible for the uptake and processing of glycerol-3-phosphate [39–41], is regulated by 21 binding

sites for five different TFs. These promoter architectures could virtually never ($P < 10^{-20}$, Eq. (2)) occur in the random promoter architecture model, and might as such be of interest for further study.

We can also use the random promoter architecture model to study the number of TF interactions per operon for particular TFs. We expect this number to be higher than suggested by the random promoter architecture model since a TF can, for example, regulate an operon cooperatively from multiple sites. As an example the well-studied Lac repressor has three known binding sites in *E. coli* [2], all regulating the same operon (*lacZYA*). Had these three sites been randomly distributed over all operons, it would have been an unlikely outcome for them all to regulate the same operon. In Table 2 we show the number of operons regulated at multiple binding sites for a given TF, both in RegulonDB 8.5 and as predicted by the random promoter architecture model [Eq. (5)]. Many of these TFs differ very significantly from the random promoter architecture model, which could be indicative of multiple TF binding domains (e.g. OxyR [42], ArgR [43]), cooperative binding (e.g. TyrR [44]), TFs which repress operons by DNA looping (e.g. NagC [45]), or chromosomal restructuring through repeated TF binding (e.g. Fis [46]). In Fig. 5 we show the specific example of OxyR regulating the *fhuF* gene at four different binding sites. Interesting exceptions include Rob and MarA, which despite being common regulators do not bind to multiple binding sites at a single operon. Thus the random promoter architecture model allows us to identify TFs of special interest.

With a large number of targets we expect global TFs to be more abundantly expressed in the cell, to avoid running the risk of depleting the reservoir of TFs and hence the TF losing its ability to function effectively [47, 48]. In Fig. 6, we explore the relationship between TF copy number and corresponding number of binding sites, using three different genome-wide protein copy number censuses based on fluorescence measurements [49], mass spectrometry [50], and ribosomal profiling [51]. We find a statistically significant positive correlation in the data set based on ribosomal profiling (log-log slope=0.61 ± 0.085), but not in the data sets based on fluorescence measurements (log-log slope=0.14 ± 0.18) or mass spectrometry (log-log slope=0.01 ± 0.17). Here, we estimate the uncertainty in the linear fit parameter using the method of bootstrapping [52]. Large systematic deviations in the protein censuses [Fig. S1] makes them difficult to use as means for model testing.

Naively one could also imagine highly expressed genes to be subject to more regulation, because expressing too many of these would be energetically costly and expressing too few could have serious consequences to the fitness of the cell. By combining binding site multiplicities from RegulonDB 8.5 with the same protein copy number censuses [49–51] we can explore the possible relationship between these two quantities. In Fig. 7 we show the number of protein copies of a gene product as a function of the number of TF binding sites regulating the gene's expression (RegulonDB 8.5). The fluorescence based census shows a weak positive relation (log-log slope=0.20 ± 0.13) between these two magnitudes, the mass spectrometry census shows no significant relation (log-log slope=0.02 ± 0.08), while the census based on ribosomal profiling presents a statistically significant negative relation (-0.17 ± 0.072). Again, the disagreement between the three protein censuses, shown in Fig. S1, makes it difficult to draw any definite conclusion regarding the relationship between gene regulation and protein expression and demonstrates a need for more rigor in the quantitative analysis of these problems.

How are activator and repressor binding sites configured?

Many genes need to be expressed only under conditions satisfying some "combinatorial rule". For example the β-galactosidase enzyme LacZ in *E. coli*, which cleaves lactose, is only highly expressed if lactose is present and glucose, the more favored energy source, is not present [53]. A general combinatorial promoter is regulated by one or more activators and repressors. Such combinatorial control requires multiple TFs to either activate and repress a gene, and the configuration of the two types of interactions determines the regulatory response. In this section we will study promoter architectures in more detail and their influence on gene expression.

To classify promoter architectures we adopt the notation (A, R) for a promoter regulated by A activator

and R repressor binding sites. Using RegulonDB we can easily find the distribution $P(A, R)$ for (A, R) with respect to all known regulatory interactions in *E. coli*. We show the most dominant promoter architectures and some specific examples in Fig. 8, along with their expected frequency in the "two-TF" random promoter architecture model described in Models [Eq. (3)]. We see an almost equal use of repressors and activators, 53 % vs. 47 % interactions, and for each promoter architecture (m, n) shown in Fig. 8(A) its symmetric counterpart (n, m) is almost equally present, both in absolute numbers and compared to the random promoter architecture model. Using the random promoter architecture model we can identify TF pairs which coregulate operons more frequently than one would expect by chance, a possible sign of TF-TF interactions or two TFs with otherwise related biological function [25]. In Table 3 we list the ten most such overrepresented TF pairs. The top pairs are all possible pairwise combinations of the MarA, SoxS and Rob TFs. These TFs are all paralogous proteins, having around 45% identical amino acid sequence at their N-terminals [54], responsible for regulating various stress responses. Note that some TFs might recognize similar or identical DNA sequences. In fact, this is the case of SoxS, Rob and MarA [55], and GalR and GalS [56]. FNR and ArcA are both global regulators responding to the availability of oxygen [57] in the cellular environment. NarL and NarP are homologous proteins responding to availability of nitrate and nitrite [58], and have been shown to act (anti)cooperatively with FNR [59, 60]. Fur and IHF are also global regulators, whose interplay with FNR has been investigated in [61–63]. GalR-GalS are homologous proteins responding to galactose [64], and GadX-GadW are homologous proteins responding to variations in pH level [37]. Even though TF pairs like Fis-CRP are more frequent coregulators (at 38 operons) in absolute numbers than any of the TF pairs listed in Table 3, this pair is not particularly overrepresented when compared to the random promoter architecture model (p-value "only" 10^{-3}), and their frequent coregulation can simply be attributed to the large number of CRP and Fis binding sites. Hence the random promoter architecture model allows us to find the most interesting TF pairs [Table 3].

Having identified the most common promoter architectures we are curious to find out how these relate to gene expression. Is there any relationship between promoter architecture and gene expression level for steady-state growth? To answer this question we identify all genes corresponding to a certain promoter architecture (A, R) in RegulonDB and acquire the protein copy number distribution of these genes from the three *E. coli* protein censuses [49–51] [see Fig. 9]. Perhaps surprisingly, we find no systematic correlation between the number of activator and repressor binding sites, and gene expression in the three sets covering thousands of genes. The only exception is the promoter architecture with one activator and one repressor binding site each (1,1), whose median expression level is higher than the upper quartile of the other five studied promoter architectures, indicating that genes with this architecture might be more abundantly expressed. Still, the figure shows that even for a given promoter architecture there is a vast spread in protein copy number, spanning up to three orders of magnitude. It seems likely that all promoter architectures in Fig. 9 would be capable of producing proteins across the full range of biologically relevant concentrations. The main purpose of activators appears not to be increasing the maximum possible expression of a gene but rather, together with repressors, modulating it around a certain mean level. This mean level can on the other hand be achieved through other mechanisms, such as the ribosomal binding sequence (RBS) or promoter strength, which we will discuss in a later section.

Where are TF binding sites located?

There are many different ways in which TFs can regulate the transcription rate of a gene. Perhaps most intuitively TFs can facilitate or block RNAP from interacting with a promoter of interest, to either activate or repress transcription of a gene [14]. However, TFs can modulate basically any step in the chain of events preceding promoter escape [65], or modify the DNA methylation or compactification states [66]. In eukaryotes, where the latter regulatory strategies are common, TF binding sites can be located hundreds of thousands of base pairs away from the transcription start site, which means that DNA needs to "loop" to establish a contact between TF and RNAP (if necessary for regulation) [66]. Hence each class of transcriptional regulation will have its own TF binding profile, and in this section we

will investigate these profiles in more detail for *E. coli*.

After aligning all known promoters with respect to their transcription start site we can make a histogram [see Fig. 10] over the number of binding sites overlapping each nucleotide position. In eukaryotes, particularly in metazoans, DNA compaction through architectural complexes such as nucleosomes can bring TF binding sites in close physical proximity to promoters located millions of base pairs downstream [67]. As bacteria lack many of these architectural complexes, we hypothesize that binding sites in bacteria are constrained to be located closer to the promoter, leading to a narrower distribution of binding sites around the promoter as compared to eukaryotes. In fact 75 % of all reported TF interactions in RegulonDB 8.5 take place within 100 bp of the transcription start site.

Activator and repressor binding sites have fundamentally different profiles; whereas repressors overlap the RNAP binding site for maximum repression, activators typically facilitate transcription initiation from upstream of the -35 region. TFs binding significantly upstream of -35 bp would, to a larger extent, need to loop DNA to interact directly with RNAP, or regulate expression of genes through other long range mechanisms. An interesting difference between specific activators [Fig. 10(C)] and global activators [Fig. 10(A)], is that the latter have two separate peaks, located at -70 bp and -45 bp respectively, rather than one. The TFs whose contribution dominates these two peaks, which should correspond to class I and class II activation [65, 68], are CRP and Fis (shown separately in Fig. 10(A)). Class I activators interact with the α-CTD domain, whereas class II activators interact directly with the sigma factor.

Although most repressors function by blocking RNAP from binding the promoter, still roughly 25% of the repressors bind upstream of -70 bp, i.e. without the possibility of blocking RNAP [69–72]. Additional mechanisms through which an upstream repressor could block transcription is by forming DNA loops to contact the transcriptional machinery as well as downstream operators [9, 73]. Another possible way these upstream repressors could function is by preventing activators from binding the promoter, or inhibiting an activator from accessing its binding site without overlapping it via DNA allostery [10] or DNA bending [74, 75]. To test this hypothesis we show in Fig. 11 the probability of a repressor binding site overlapping an activator binding site as a function of position, using the probability for two activators to overlap as comparison. The results show that around 30 % of the repressors binding upstream of -70 bp overlap with an activator, compared to 15-20 % for two different activators. This suggests that blocking of activators is an important regulatory strategy for upstream repressors but not the only one, as a large fraction of upstream repressors inhibit transcription through other means.

In total, almost half of all binding sites reported in RegulonDB 8.5 overlap with other binding sites, which leads us to believe that this constitutes an important regulatory strategy. As more binding sites are discovered, the number of overlapping binding sites will likely increase, just as the probability of two students in a class having birthday on the same day goes up rapidly with the number of students. Interestingly, TFs often (37 % of the reported overlapping interactions) overlap with themselves. For example, out of the 88 known Fur binding sites, 75% of them are reported to overlap with other Fur binding sites [76].

Since the regulatory region of a gene is of limited size, TFs need to compete for space at promoters with other binding sites, in particular TFs which interact directly with RNAP. To study this "real estate" problem we first collect the DNA binding site size of all TF-DNA interaction sites reported in RegulonDB 8.5 [see Fig. 12]. A similar figure is reported in [77] using an earlier version of Regulon DB. Some of the notable peaks in Fig. 12 correspond mainly to binding of global TFs: Fis (15 bp), ArcA (15 bp) and CRP (22 bp). Most bacterial TFs interact with DNA along a contiguous region of around 15 bp (although outliers exist) which means that one could theoretically fit three nonoverlapping binding sites within 50 bp. Since the majority of operons reportedly have fewer than this number of binding sites [see Fig. 3], the size of the regulatory region does not in general seem to be a major constraining factor. However, for promoters with a larger number of binding sites, of which we saw some examples in Fig. 3, TFs would either need to bend DNA to access RNAP, or overlap with other TFs. To further study the real estate of the promoter we look at the separation between binding sites [see Fig. 13], which shows the

the edge-to-edge distance for nonoverlapping adjacent binding sites. The majority of binding sites in this set are separated by less than 15 bp from their neighbors. Hence for an operon with three binding sites the regulatory region would be expected to take up around $3 \times 15 + 2 \times 15 = 75$ bp, around the same as observed in Fig. 10.

How does promoter architecture relate to promoter strength?

Many prokaryotic genes do not rely on TFs for regulation, and will be constitutively expressed independently of the cellular environment. The production of these genes will, at our current best understanding, only be affected by the global availability of RNAP, sigma factors, ribosomes and the interaction strengths with these different complexes [78]. For proteins which are in constant demand, constitutive expression provides a simple and efficient choice of promoter architecture. Despite its simplicity, constitutive expression allows for an impressive dynamic range in protein production [26], as is also suggested by Fig. 9. This demonstrates the power of the basal production machinery, whose transcriptional component we will study further in this section. In particular we will be interested in the relationship between basal promoter strength and regulation by TFs.

In *E. coli* the transcription rate of a gene can vary by up to three orders of magnitude due to differences in the promoter strength alone [26], not taking TFs into account. To illustrate this point we use the linear RNAP-DNA interaction model introduced in Models to predict the binding energy to all known σ^{70} promoters along with the corresponding distribution for nonspecific binding [see Fig. 14]. As expected we get two separate distributions, where RNAP binds on average 2.4 k_BT more strongly to known promoters than sequences chosen randomly from the *E. coli* genome. The predicted RNAP binding energy distribution spans roughly 8 k_BT from the strongest to the weakest promoter, corresponding to a predicted 3000-fold difference in RNAP binding affinity. This difference is similar to that found between the most abundantly expressed TFs (e.g. CRP) and scarcely expressed TFs (e.g. LacI) [79, 80] in *E. coli*, suggesting that promoter strength alone might be a powerful enough tool to set the mean level of gene expression to most biologically relevant values. Analysis of promoter sequences has revealed that functional transcriptional start sites are surrounded by noninitiating pseudopromoters [86]. A discomforting observation from Fig. 14, however, is that 200,000 sites or so in the 5×10^6 bp *E. coli* background interact more strongly with RNAP than the typical promoter. This raises several important questions [81–83]: Is the linear energy model missing key information, or can all the predicted promoters in principle produce transcripts? Do weak promoters need to be activated by TFs to function? Although trying to solve these important questions falls outside the scope of the current paper, we note that the paradox might originate from the fact that the promoter sequence encodes detailed information about both RNAP binding affinity, open complex formation rate and promoter escape rate [84, 85] in a way that likely cannot be captured in a simple linear model. Powerful new methods such as RNA-seq [87] could provide further insight into which of the 200,000 predicted promoters are actually transcriptionally active.

In Fig. 9 we learned that the number of activator or repressor binding sites did not seem to have any systematic effect on the average gene expression in three sets covering thousands of genes. Since activators, by definition, increase the expression of a gene and repressors reduce it, the only possible explanation for this observation (if true) is that repressed genes have a higher basal level of expression. This could, for example, be the result if repressed genes have a higher affinity (*promoter strength*) for RNAP to their promoters. Since stronger promoters recruit RNAP more easily they would hence become transcriptionally more active. To investigate the relationship between promoter strength and promoter architecture we show in Fig. 15 the RNAP binding energy distribution for genes which according to RegulonDB 8.5 are regulated from a single activator or repressor binding site. For these promoters our data suggest, though not conclusively, that RNAP binds more strongly to the promoters of repressed genes, $\bar{E}_{\text{repressed}} = -2.0 \pm 0.17\ k_BT$, than promoters of activated genes, $\bar{E}_{\text{activated}} = -1.1 \pm 0.23\ k_BT$. The reported uncertainty in the means are estimated using the method of bootstrapping [52]. Our results suggest that repressed genes have a higher basal rate of transcription, providing a possible explanation as

to why we do not see a significant difference in gene expression as compared to activated genes. Conversely, weak promoters are more likely to be activated by TFs, suggesting that these promoters might not work effectively without TF activation.

Discussion

After more than half a century of intense study *E. coli* remains one of the most important model organisms in biology. In order to make the vast pool of knowledge obtained from these studies publicly available and directly accessible, ambitious initiatives such as RegulonDB have curated thousands of references to collect information relating to TF binding site locations, organization of transcriptional units, and more. Although this annotation process is far from complete, as more than half of the *E. coli* genes still lack any known regulation, we now have a better opportunity than ever to study regulatory interactions in detail.

In this study we have analyzed TF-DNA interactions reported in RegulonDB 8.5. We find distinct differences in binding site location trends depending on TF type: activator or repressor, global or specific TFs. To study promoter architectures in greater depth we created a random promoter architecture model. This random model makes it possible to generate "null hypotheses" for promoter architectures which can then be compared to real regulatory architectures from RegulonDB data. Our findings can be summarized as follows:

1. We find that most promoters in *E. coli* are less heavily regulated than expected from the random promoter architecture model. The majority of operons in RegulonDB 8.5 have fewer than three known associated TF binding sites, and most specific TFs regulate fewer than three operons, suggesting that many *E. coli* genes are expressed with little "oversight" from TFs. Some interesting exceptions include operons such as gadAXW (acid resistance), csgDEFG (Curli amyloid fibers), and glpTQ (glycerol-3-phosphate uptake) which are controlled by up to 30 binding sites.

2. The random promoter architecture model allows us to identify, with well defined statistical significance, pairs of TFs which frequently coregulate operons, e.g. due to cooperative interactions or recognition of similar consensus sequences. Examples include the stress regulators MarA/Rob/SoxS, and the oxygen responding TFs FNR/ArcA. The random model further allows us to recognize TFs such as OxyR and Fis, which frequently bind to multiple binding sites per operon, e.g. due to cooperative binding, DNA looping, or through multiple binding domains. Our method of comparing promoter architectures to a null hypothesis provides a new approach for detecting coregulation and allows us to formulate experimentally testable hypotheses using only a list of known TF binding sites regulating each operon.

3. We find no systematic correlation between the number of activating or repressing TF interactions and the mean expression of a gene, as measured by three different genome-wide protein censuses covering thousands of genes. A position-weight-matrix model used to estimate the binding affinity of RNAP to promoters of activated and repressed genes, suggested that this lack of correlation might in part be due to differences in basal transcription rates of promoters. In this scenario, promoters that are being repressed have a higher basal level than promoters that are being activated.

One of the grand challenges of physical biology is to be able to construct predictive maps between promoter nucleotide sequence and gene expression. Increasingly accurate promoter architecture data, found e.g. using powerful techniques like ChIP-Seq, allow predictive maps to be both tested and refined. A difficulty with mapping promoter architecture and gene expression, apart from lacking complete knowledge of the regulatory network, is a substantial disagreement on protein concentrations as measured using different experimental methods and under different experimental conditions. The protein copy numbers measured using mass spectrometry [50] are for example on average at least one order of magnitude higher than for the same proteins measured with fluorescence based techniques [49], though these kinds of effects

can be due to different growth conditions for the cells [78, 88]. As TF copy number plays a central role in regulatory function, we believe resolving these discrepancies will be a necessary step for a deeper understanding of several important aspects of gene regulation. To become quantitatively predictive, gene regulatory maps must come to relate gene expression data with precise promoter architecture data such as binding site locations and binding energies. These will allow for an accurate *in silico* description of global promoter activity, and provide quantitative predictions for genome-scale experiments.

Acknowledgments

We would like to thank Genya Frenkel and Ron Milo for comments on early versions of the manuscript.

References

1. Garcia HG, Phillips R (2011) Quantitative dissection of the simple repression input-output function. Proc Natl Acad Sci U S A 108: 12173-8.

2. Oehler S, Eismann ER, Kramer H, Muller-Hill B (1990) The three operators of the *lac* operon cooperate in repression. EMBO J 9: 973-9.

3. Segal E, Widom J (2009) From DNA sequence to transcriptional behaviour: A quantitative approach. Nat Rev Genet 10: 443-56.

4. Zentner GE, Henikoff S (2013) Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol 20: 259–266.

5. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, et al. (2013) RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 41: D203–213.

6. Keseler IM, Collado-Vides J, Santos-Zavaleta A, Peralta-Gil M, Gama-Castro S, et al. (2011) EcoCyc: A comprehensive database of *Escherichia coli* biology. Nucleic Acids Res 39: D583–590.

7. Ackers GK, Johnson AD, Shea MA (1982) Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci U S A 79: 1129-33.

8. Müller J, Oehler S, Müller-Hill B (1996) Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator. J Mol Biol 257: 21-9.

9. Garcia HG, Sanchez A, Boedicker JQ, Osborne M, Gelles J, et al. (2012) Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. Cell Rep 2: 150–161.

10. Kim S, Broströmer E, Xing D, Jin J, Chong S, et al. (2013) Probing allostery through DNA. Science 339: 816–819.

11. Ryu S, Fujita N, Ishihama A, Adhya S (1998) GalR-mediated repression and activation of hybrid lacUV5 promoter: differential contacts with RNA polymerase. Gene 223: 235-45.

12. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. Science 316: 1497-502.

13. Kinney JB, Murugan A, Callan CG, Cox EC (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci USA 107: 9158–9163.

14. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. (2005) Transcriptional regulation by the numbers: Models. Curr Opin Genet Dev 15: 116-24.

15. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, et al. (2005) Transcriptional regulation by the numbers: Applications. Curr Opin Genet Dev 15: 125-35.

16. Sherman MS, Cohen BA (2012) Thermodynamic state ensemble models of cis-regulation. PLoS Computational Biology 8: e1002407.

17. Buchler NE, Gerland U, Hwa T (2003) On schemes of combinatorial transcription logic. Proc Natl Acad Sci U S A 100: 5136-41.

18. Erdos P, Renyi A (1959) On random graphs i. Publ Math Debrecen 6: 290–297.

19. Alon U (2006) An Introduction to Systems Biology: Design Principles of Biological Circuits (Chapman & Hall/CRC Mathematical & Computational Biology). Chapman and Hall/CRC, First edition.

20. Madan Babu M, Teichmann SA (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res 31: 1234–1244.

21. Ali Azam T, Iwata A, Nishimura A, Ueda S, Ishihama A (1999) Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. J Bacteriol 181: 6361–6370.

22. Abramowitz M, Stegun IA (1964) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, Ninth edition.

23. Nishimura K, Sibuya M (1988) Occupancy with two types of balls. Annals of the Institute of Statistical Mathematics 40: 77-91.

24. Reif F (1965) Fundamentals of statistical and thermal physics. New York,: McGraw-Hill.

25. Ho Sui SJ, Fulton DL, Arenillas DJ, Kwon AT, Wasserman WW (2007) oPOSSUM: Integrated tools for analysis of regulatory motif over-representation. Nucleic Acids Res 35: W245–252.

26. Brewster RC, Jones DL, Phillips R (2012) Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*. PLoS Comput Biol 8: e1002811.

27. Mulligan ME, Hawley DK, Entriken R, McClure WR (1984) *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity. Nucleic Acids Research 12: 789–800.

28. Brunner M, Bujard H (1987) Promoter recognition and promoter strength in the *Escherichia coli* system. Embo J 6: 3139-44.

29. Stormo GD (2000) DNA binding sites: Representation and discovery. Bioinformatics 16: 16–23.

30. Gross CA, Chan C, Dombroski A, Gruber T, Sharp M, et al. (1998) The functional and regulatory roles of sigma factors in transcription. Cold Spring Harb Symp Quant Biol 63: 141–155.

31. Macnab RM (1992) Genetics and biogenesis of bacterial flagella. Annu Rev Genet 26: 131–158.

32. Mermod M, Magnani D, Solioz M, Stoyanov JV (2012) The copper-inducible ComR (YcfQ) repressor regulates expression of ComC (YcfR), which affects copper permeability of the outer membrane of *Escherichia coli*. Biometals 25: 33–43.

33. Hirashima A, Wang S, Inouye M (1974) Cell-free synthesis of a specific lipoprotein of the *Escherichia coli* outer membrane directed by purified messenger RNA. Proc Natl Acad Sci USA 71: 4149–4153.

34. Takahashi S, Hours C, Chu A, Denhardt DT (1979) The rep mutation. VI. Purification and properties of the *Escherichia coli* rep protein, DNA helicase III. Can J Biochem 57: 855–866.

35. Laimins LA, Rhoads DB, Epstein W (1981) Osmotic control of kdp operon expression in *Escherichia coli*. Proc Natl Acad Sci USA 78: 464–468.

36. Thelen P, Tsuchiya T, Goldberg EB (1991) Characterization and mapping of a major Na+/H+ antiporter gene of *Escherichia coli*. J Bacteriol 173: 6553–6557.

37. Tramonti A, De Canio M, De Biase D (2008) GadX/GadW-dependent regulation of the *Escherichia coli* acid fitness island: Transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. Mol Microbiol 70: 965–982.

38. Hammar M, Arnqvist A, Bian Z, Olsen A, Normark S (1995) Expression of two csg operons is required for production of fibronectin- and congo red-binding curli polymers in *Escherichia coli* K-12. Mol Microbiol 18: 661–670.

39. Wanner BL (1990) Phosphorus Assimilation and Its Control of Gene-Expression in Escherichia-Coli. Biological Chemistry Hoppe-Seyler 371: 180-180.

40. Rao NN, Roberts MF, Torriani A, Yashphe J (1993) Effect of glpT and glpD mutations on expression of the phoA gene in *Escherichia coli*. J Bacteriol 175: 74–79.

41. Vershinina OA, Znamenskaya LV (2002) The Pho regulons of bacteria. Microbiology 71: 497-511.

42. Toledano MB, Kullik I, Trinh F, Baird PT, Schneider TD, et al. (1994) Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: A mechanism for differential promoter selection. Cell 78: 897–909.

43. Tian G, Lim D, Carey J, Maas WK (1992) Binding of the arginine repressor of *Escherichia coli* K12 to its operator sites. J Mol Biol 226: 387–397.

44. Bai Q, Somerville RL (1998) Integration host factor and cyclic AMP receptor protein are required for TyrR-mediated activation of *tpl* in *Citrobacter freundii*. J Bacteriol 180: 6173–6186.

45. Plumbridge J, Kolb A (1993) DNA loop formation between Nag repressor molecules bound to its two operator sites is necessary for repression of the nag regulon of *Escherichia coli* in vivo. Mol Microbiol 10: 973–981.

46. Schneider R, Lurz R, Luder G, Tolksdorf C, Travers A, et al. (2001) An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. Nucleic Acids Res 29: 5107–5114.

47. Rydenfelt M, Cox RS, Garcia H, Phillips R (2014) Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. Phys Rev E 89: 012702.

48. Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, et al. (2014) The transcription factor titration effect dictates level of gene expression. Cell 156: 1312–1323.

49. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, et al. (2010) Quantifying *E. coli* proteome and rranscriptome with single-molecule sensitivity in single cells. Science 329: 533–538.

50. Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol 25: 117-24.

51. Li GW, Burkhardt D, Gross C, Weissman JS (2014) Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624-35.

52. Efron B (1979) Bootstrap methods: Another look at the jackknife. Ann Statistics 7: 1-26.

53. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3: 318–356.

54. Cohen SP, Hachler H, Levy SB (1993) Genetic and functional analysis of the multiple antibiotic resistance (*mar*) locus in *Escherichia coli*. J Bacteriol 175: 1484–1492.

55. Martin RG, Gillette WK, Rhee S, Rosner JL (1999) Structural requirements for marbox function in transcriptional activation of mar/sox/rob regulon promoters in Escherichia coli: sequence, orientation and spatial relationship to the core promoter. Mol Microbiol 34: 431–441.

56. Geanacopoulos M, Adhya S (1997) Functional characterization of roles of GalR and GalS as regulators of the gal regulon. J Bacteriol 179: 228-34.

57. Levanon SS, San KY, Bennett GN (2005) Effect of oxygen on the *Escherichia coli* ArcA and FNR regulation systems and metabolic responses. Biotechnol Bioeng 89: 556–564.

58. Rabin RS, Stewart V (1993) Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in *Escherichia coli* K-12. J Bacteriol 175: 3259–3268.

59. Darwin AJ, Ziegelhoffer EC, Kiley PJ, Stewart V (1998) Fnr, NarP, and NarL regulation of *Escherichia coli* K-12 *napF* (periplasmic nitrate reductase) operon transcription in vitro. J Bacteriol 180: 4192–4198.

60. Overton TW, Griffiths L, Patel MD, Hobman JL, Penn CW, et al. (2006) Microarray analysis of gene regulation by oxygen, nitrate, nitrite, FNR, NarL and NarP during anaerobic growth of *Escherichia coli*: New insights into microbial physiology. Biochem Soc Trans 34: 104–107.

61. Myers KS, Yan H, Ong IM, Chung D, Liang K, et al. (2013) Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. PLoS Genet 9: e1003565.

62. Mettert EL, Kiley PJ (2007) Contributions of [4Fe-4S]-FNR and integration host factor to fnr transcriptional regulation. J Bacteriol 189: 3036–3043.

63. Troxell B, Fink RC, Porwollik S, McClelland M, Hassan HM (2011) The Fur regulon in anaerobically grown *Salmonella enterica* sv. Typhimurium: Identification of new Fur targets. BMC Microbiol 11: 236.

64. Weickert MJ, Adhya S (1992) Isorepressor of the gal regulon in *Escherichia coli*. J Mol Biol 226: 69–83.

65. Dove SL, Joung JK, Hochschild A (1997) Activation of prokaryotic transcription through arbitrary protein-protein contacts. Nature 386: 627–630.

66. Ptashne M, Gann A (2002) Genes and Signals. New York: Cold Spring Harbor Laboratory Press.

67. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.

68. Busby S, Ebright RH (1999) Transcription activation by catabolite activator protein (CAP). J Mol Biol 293: 199-213.

69. Hochschild A, Dove SL (1998) Protein-protein contacts that activate and repress prokaryotic transcription. Cell 92: 597–600.

70. Pavco PA, Steege DA (1990) Elongation by Escherichia coli RNA polymerase is blocked in vitro by a site-specific DNA binding protein. J Biol Chem 265: 9960–9969.

71. Pavco PA, Steege DA (1991) Characterization of elongating T7 and SP6 RNA polymerases and their response to a roadblock generated by a site-specific DNA binding protein. Nucleic Acids Res 19: 4639–4646.

72. Rojo F (1999) Repression of transcription initiation in bacteria. J Bacteriol 181: 2987–2991.

73. Garcia HG, Grayson P, Han L, Inamdar M, Kondev J, et al. (2007) Biological consequences of tightly bent DNA: The other life of a macromolecular celebrity. Biopolymers 85: 115-30.

74. Pérez-Martín J, Rojo F, De Lorenzo V (1994) Promoters responsive to DNA bending: A common theme in prokaryotic gene expression. Microbiological Reviews 58: 268–290.

75. Kim J, Zwieb C, Wu C, Adhya S (1989) Bending of DNA by gene-regulatory proteins: Construction and use of a DNA bending vector. Gene 85: 15–23.

76. Chen Z, Lewis KA, Shultzaberger RK, Lyakhov IG, Zheng M, et al. (2007) Discovery of Fur binding site clusters in *Escherichia coli* by information theory models. Nucleic Acids Res 35: 6762–6777.

77. Ruths T, Nakhleh L (2013) Neutral forces acting on intragenomic variability shape the Escherichia coli regulatory network topology. Proc Natl Acad Sci USA 110: 7754–7759.

78. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. Science 330: 1099-102.

79. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, et al. (2008) Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 9: 102.

80. Gilbert W, Muller-Hill B (1966) Isolation of the Lac Repressor. Proc Natl Acad Sci U S A 56: 1891-1898.

81. Djordjevic M (2013) Efficient transcription initiation in bacteria: An interplay of protein-DNA interaction parameters. Integr Biol (Camb) 5: 796–806.

82. Gershenzon NI, Stormo GD, Ioshikhes IP (2005) Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Res 33: 2290–2301.

83. Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. Genome Res 13: 2381–2390.

84. McClure WR (1980) Rate-limiting steps in RNA chain initiation. Proc Natl Acad Sci USA 77: 5634–5638.

85. McClure WR (1985) Mechanism and control of transcription initiation in prokaryotes. Annu Rev Biochem 54: 171–204.

86. Huerta AM, Collado-Vides J (2003) Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals. J Mol Biol 333: 261-78.

87. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644–652.

88. Klumpp S, Zhang Z, Hwa T (2009) Growth-rate dependent global effects on gene expression in bacteria. Cell .

89. Cox RS, Surette MG, Elowitz MB (2007) Programming gene expression with combinatorial promoters. Mol Syst Biol 3: 145.

90. Martinez-Antonio A, Collado-Vides J (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. Curr Opin Microbiol 6: 482-9.

91. Seshasayee AS, Fraser GM, Babu MM, Luscombe NM (2009) Principles of transcriptional regulation and evolution of the metabolic system in E. coli. Genome Res 19: 79-91.

Figure Legends

Figure 1. Combinatorics for distribution of binding sites across the genome. (A) M operons with at least two binding sites, and k operons with exactly one binding site. The remaining $N_{op} - M - k$ operons are empty. (B) M operons with at least one binding site of each kind, and k operons with at least one binding site of first kind but none of the second. The remaining $N_{op}Mk$ operons are either empty or have only binding sites of the second kind.

Figure 2. Number of operons, genes and binding sites regulated per TF (RegulonDB 8.5.) (A) Schematic of operons regulated by the FlhCD TFs according to RegulonDB 8.5. (B) The TFs have been sorted by increasing number of interactions, and the dark shaded area highlights the TFs responsible for 50% of all regulatory interactions in *E. coli*, which we denote as global TFs. The median number of operons, genes (coding sequences) and binding sites regulated per TF is 3, 4 and 6.5, respectively. The number of regulated genes is calculated by taking into account how many coding sequences are contained within each operon.

Figure 3. Number of binding sites and TF types regulating each operon (RegulonDB 8.5). The mean number of binding sites per operon is 1.1 (3.5 for operons with at least one known binding site). The best exponential fits (in log-space) are shown in the figure as dashed lines. These fits are expected to change as more binding sites are discovered.

Figure 4. Number of TF binding sites per operon. (A) Distribution of number of TF binding sites per operon in RegulonDB 8.5 and the random promoter architecture model. Shown separately are distributions after excluding unregulated operons ("regulated only"). (B) The *glpTQ* operon is regulated by 21 binding sites for five different TFs.

Figure 5. *fhuF* regulation. The TF OxyR is enriched for regulating genes at multiple binding sites as shown in Table 2. For example *fhuF* is regulated by four binding sites for OxyR.

Figure 6. TF copy number plotted as a function of the total number of TF binding sites (RegulonDB 8.5) for that particular TF. The TF copy number is measured in [49–51]. The two lines mark the critical boundary where the number of binding sites is large enough to deplete TFs binding as monomers (solid) or dimers (dashed). Updated version of figure published in [47].

Figure 7. Measured protein copy number vs. number of TF binding sites regulating the transcription of the protein. The boxes show median, upper and lower quartiles, and the dashed lines show the range of the data. Protein data based on (A) fluorescence measurements [49], (B) mass spectrometry [50], and (C) ribosomal profiling [51].

Figure 8. Frequency of promoter architectures. (A) Frequency of the most dominant promoter architectures listed in RegulonDB 8.5, and their corresponding frequency in the random promoter architecture model. Binding site configurations with A activator and R repressor binding sites are denoted by (A, R). (B) Examples of some of the architectures featured in (A).

Figure 9. Protein copy number as a function of promoter architecture for the most common architectures. The notation (A, R) represents a promoter with A activator and R repressor binding sites. Protein data based on (A) fluorescence measurements [49], (B) mass spectrometry [50], and (C) ribosomal profiling [51].

Figure 10. Distribution of activating and repressing binding sites bound by global TFs and specific TFs, respectively. The y-axis shows number of binding sites overlapping each nucleotide position, after aligning all promoters with respect to their transcription start site (TSS) for the different kinds of TFs. Similar figures were reported in [20] and [89] using earlier versions of RegulonDB.

Figure 11. Probability of TF binding site overlap. Binding sites are defined as an interval of nucleotides from the 5×10^6 bp *E. coli* genome covered by a TF upon binding. Two binding sites sharing one or more nucleotides are considered to be overlapping, independently of which strand the TFs bind. Binding sites overlapping more than one binding site are classified according to the site with the most overlap. Notice that the probabilities of "Repressor overlaps activators" and "Activator overlaps repressor" are not identical despite the number of overlapping activator and repressor binding sites in a region being fixed. For example, there are many more activators than repressors binding upstream of -100 bp, which results in a higher probability for a repressor to overlap with an activator in this region than vice versa.

Figure 12. DNA binding site size (in base pairs) for all TF-DNA interactions (RegulonDB 8.5). Mean DNA binding site size size: 17.3 bp. Also see figure published in [77].

Figure 13. Edge to edge distance between adjacent binding sites (RegulonDB 8.5). Figure does not include binding sites separated by more than 150 bp, which would likely correspond to regulation of different operons.

Figure 14. Predicted RNAP binding energy [26] for promoters in RegulonDB 8.5 and DNA sequences randomly chosen from the *E. coli* genome. The spacer region is allowed to range from 16-18 bp, and the -10 box is allowed to deviate by one base pair upstream or downstream from its consensus position. The RNAP binding energy is taken as the minimum binding energy of these $3 \times 3 = 9$ possible binding configurations.

Figure 15. Predicted RNAP binding energy to promoters in the simple activation (1,0) and simple repression architecture (0,1). These calculations represent the basal state in which no TFs are present. Operons whose transcription is initiated from multiple promoters are excluded.

Tables

TF	Operons	Genes	Binding sites
CRP	221	495	320
FNR	108	296	131
Fis	96	225	237
IHF	76	219	114
H-NS	70	179	105
ArcA	64	172	118
Fur	63	129	122
Lrp	41	103	103

Table 1. Global TFs and their associated number of binding sites, the number of operons regulated, and the total number of genes (coding sequences) regulated by each TF (RegulonDB 8.5). See Supporting Information [Table S1] for a corresponding table that includes specific TFs. The notion of global TF is not unambiguously defined, and the list presented here might therefore differ slightly from that used in other works [90, 91].

TF	Total number of binding sites	Operons regulated by multiple binding sites (RegulonDB)	Operons regulated by multiple binding sites (random promoter)	*p*-value
OxyR	44	19	0.69	1.9×10^{-31}
ArgR	34	15	0.41	3.4×10^{-27}
NarP	21	10	0.16	7.7×10^{-22}
NarL	98	25	3.3	1.4×10^{-19}
Fis	237	52	18	2.7×10^{-17}
TyrR	19	8	0.13	2.0×10^{-16}
FlhDC	30	10	0.32	1.0×10^{-15}
IHF	114	25	4.5	1.5×10^{-15}
CRP	320	67	31	3.5×10^{-14}
CytR	23	8	0.19	4.8×10^{-14}
NagC	23	8	0.19	4.8×10^{-14}

Table 2. TFs which are significantly enriched for multiple binding sites per operon, compared to the random promoter architecture model. The p-value for data in RegulonDB 8.5 is given by the probability of an equal or more extreme outcome in the random promoter architecture model. The particular example of OxyR regulating the *fhuF* at four binding sites is shown in Fig. 5. An extended version of Table 2, covering 115 TFs, is available in Supporting Information [Table S2].

TF 1	**TF 2**	**Total binding sites (TF 1)**	**Total binding sites (TF 2)**	**Coregulated operons (RegulonDB)**	**Coregulated operons (random promoter)**	***p*-value**
MarA	SoxS	24	29	18	0.26	5.5×10^{-34}
MarA	Rob	24	17	14	0.15	1.2×10^{-28}
SoxS	Rob	29	17	14	0.18	4.6×10^{-27}
FNR	ArcA	131	118	30	5.3	6.4×10^{-16}
FNR	NarL	131	98	27	4.5	3.0×10^{-15}
NarP	NarL	21	98	11	0.75	1.7×10^{-11}
FNR	Fur	131	122	24	5.5	2.8×10^{-10}
FNR	IHF	131	114	23	5.2	4.3×10^{-10}
GalR	GalS	12	12	5	0.054	5.5×10^{-10}
GadX	GadW	37	20	6	0.27	1.5×10^{-7}

Table 3. TF pairs which show significant enrichment for coregulation of operons, compared to the random promoter architecture model. The p-value for data in RegulonDB 8.5 is given by the probability of an equal or more extreme outcome in the random promoter architecture model. An extended version of Table 3, covering over 900 TF pairs, is available in Supporting Information [Table S3].

Supporting Information Legends

Figure S1. Comparison of different *E. coli* protein censuses. Measured protein copy number using mass spectrometry [50], fluorescence [49], and ribosomal profiling [51]. Note how all measurements show a systematic deviation with respect to each other. This deviation can be up to two orders of magnitude, corresponding to comparing mass spectrometry and fluorescence.

Figure S2. Number of binding sites and TF types regulating each operon (RegulonDB 8.5) shown separately for global TF binding sites (black) and specific TF binding sites (red).

Figure S3. Number of TF binding sites per operon. Distribution of number of TF binding sites per operon in RegulonDB 8.5 for (A) Global TF binding site and (B) Specific TF binding sites. Shown separately are distributions after excluding unregulated operons ("regulated only").

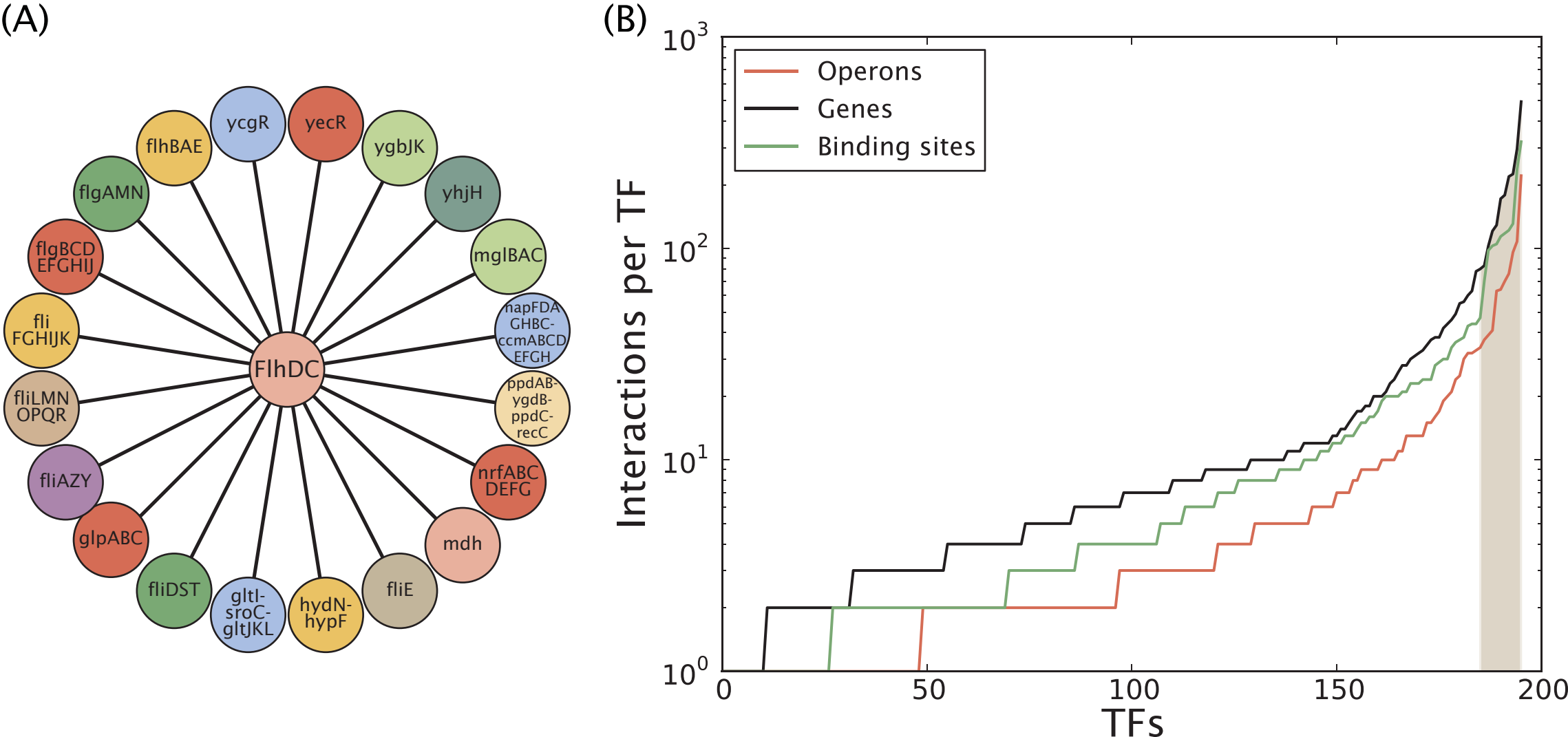
Table S1. All TFs and their associated number of binding sites, the number of operons regulated, and the total number of genes (coding sequences) regulated by each TF (RegulonDB 8.5).

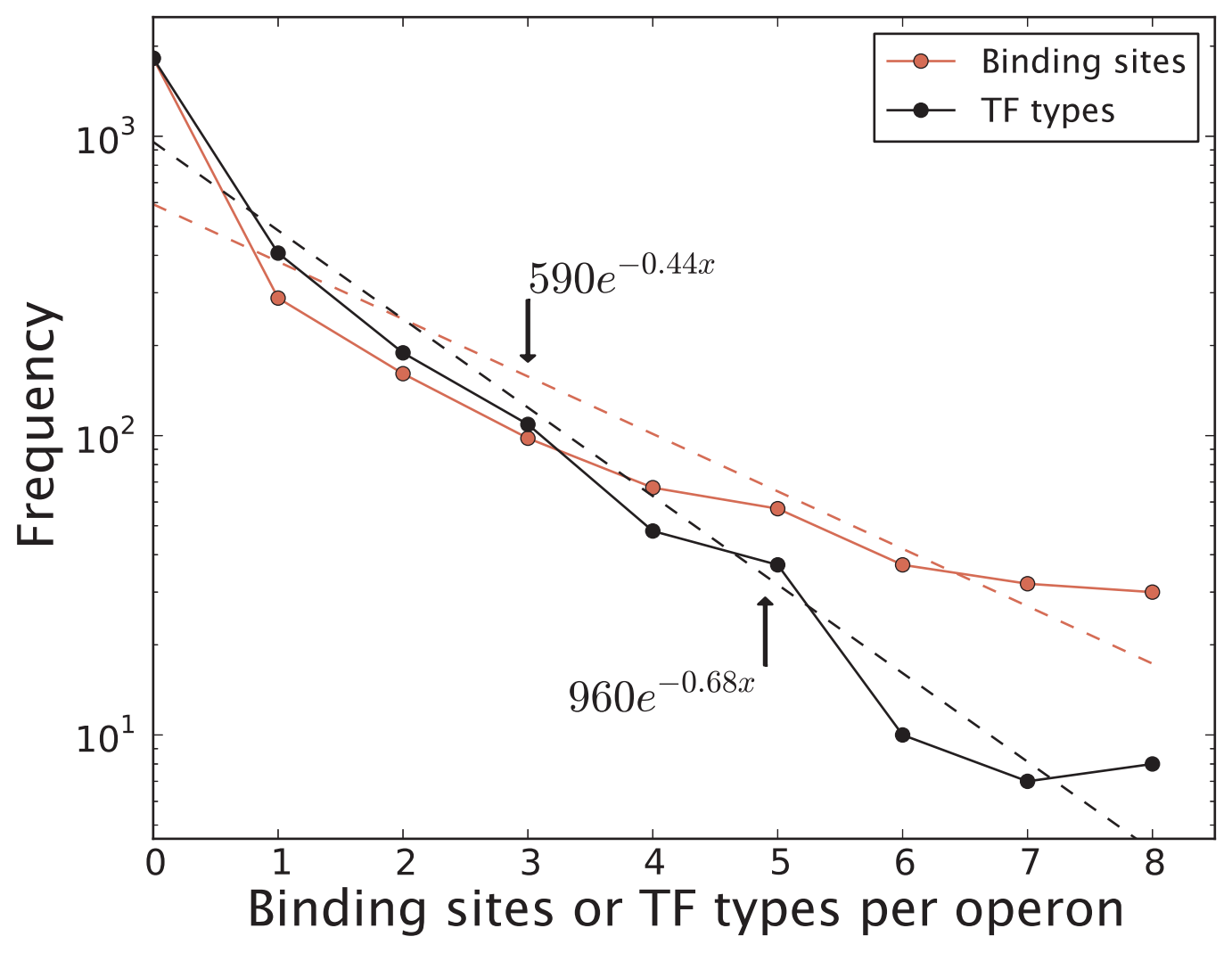
Table S2. TFs which are significantly enriched for multiple binding sites per operon, compared to the random promoter architecture model.

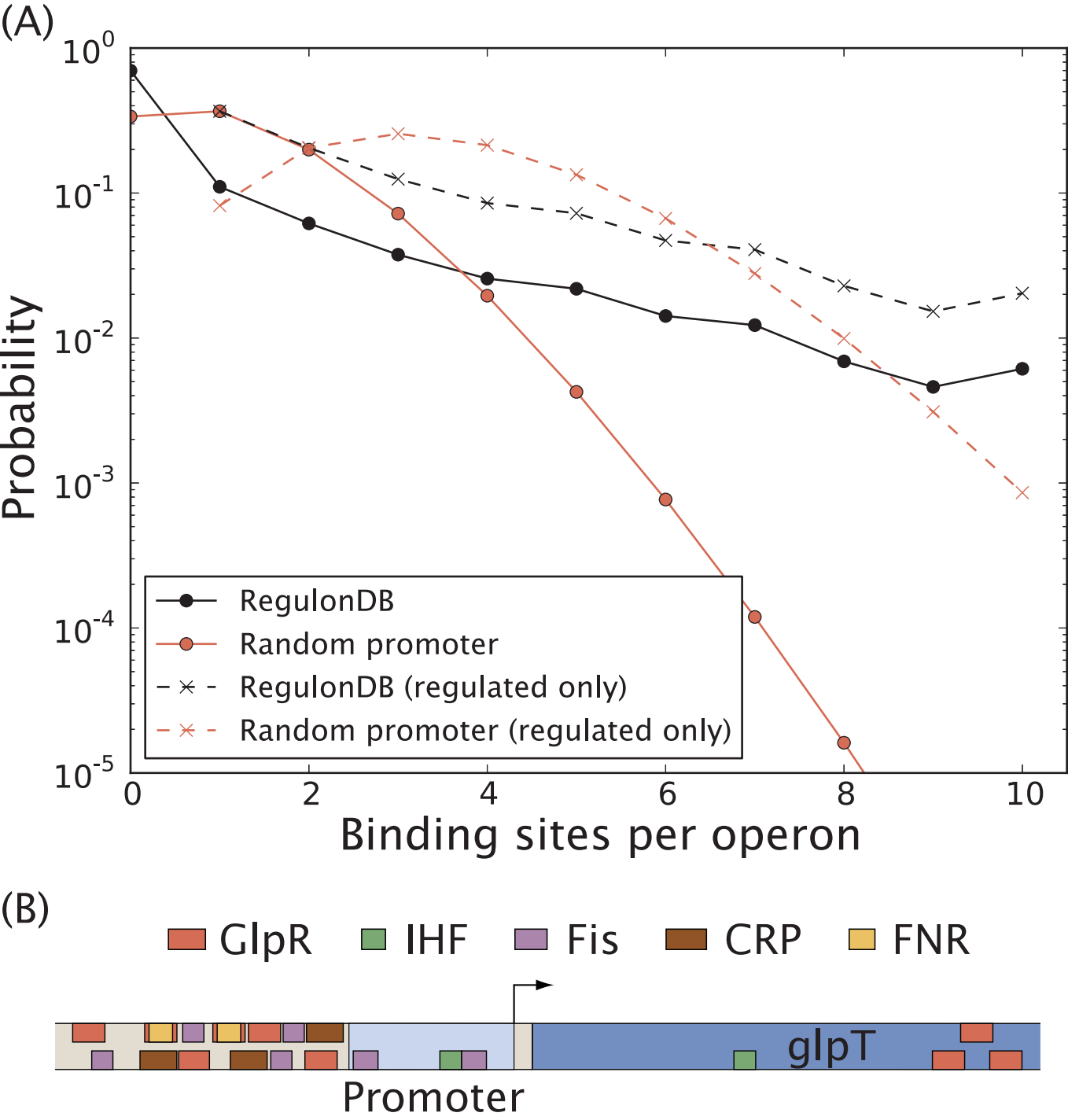
Table S3. TF pairs which show significant enrichment for coregulation of operons, compared to the random promoter architecture model.

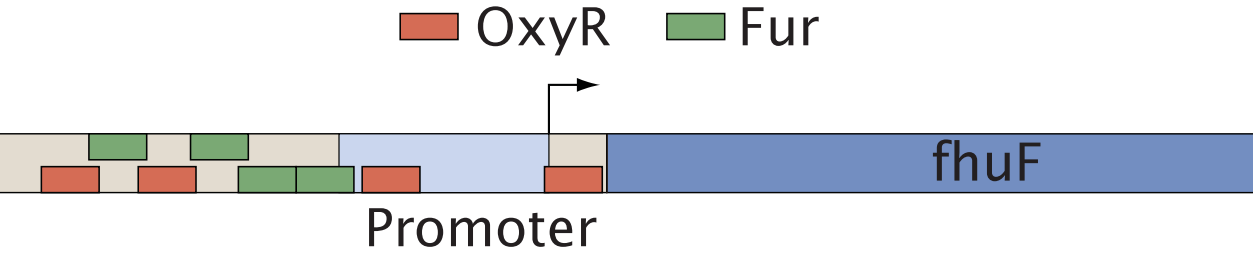
(A) One TF type
N_{op} - M - k operons with no binding sites
k operons with 1 binding site
M operons with 2 or more binding sites
Binding sites
Promoter
N_{op} operons

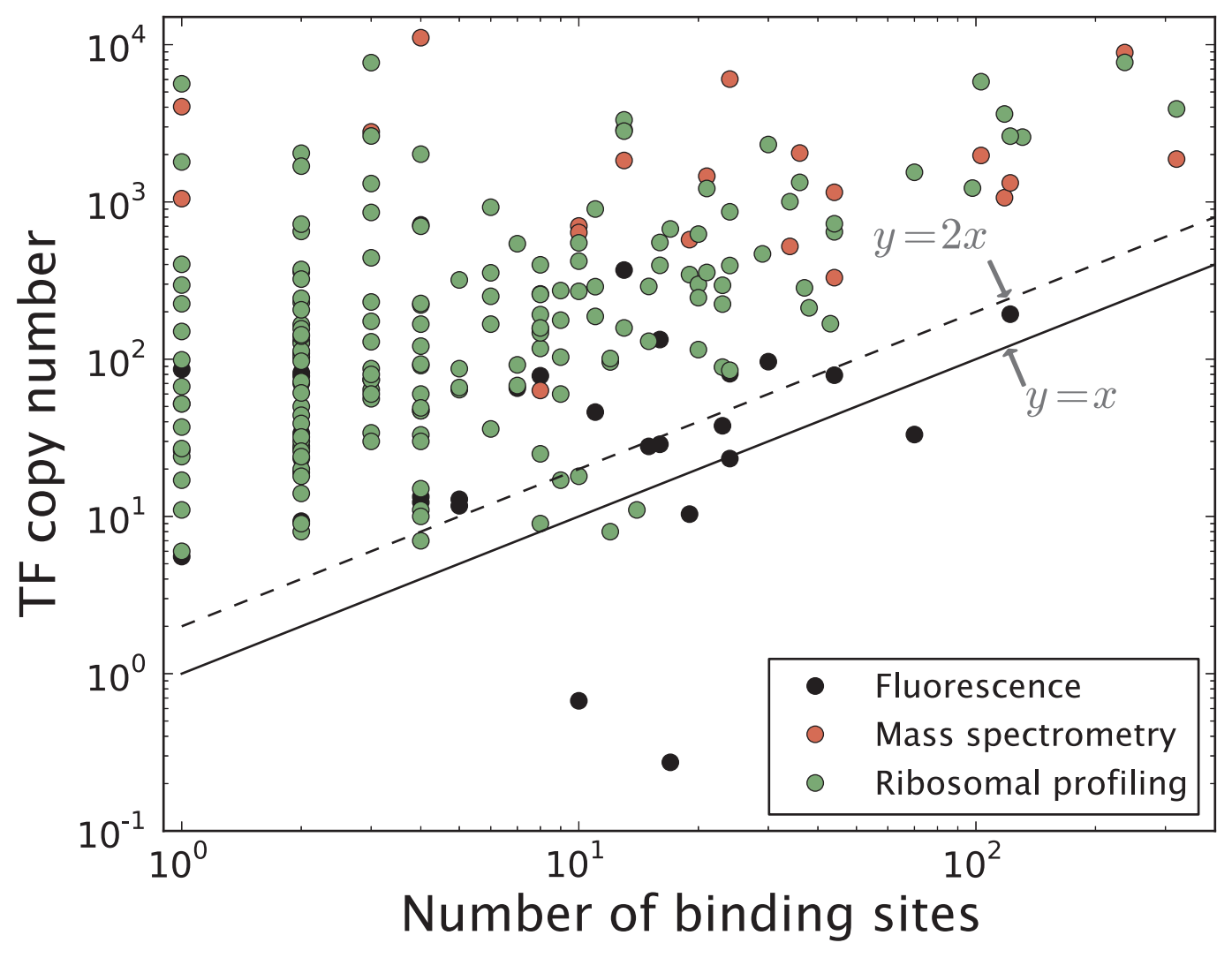
(B) Two TF types
N_{op} - M - k operons with no binding sites or binding sites of a second kind only
k operons with only binding site 1
M operons with at least one binidng site of each kind
Binding sites
Promoter
N_{op} operons

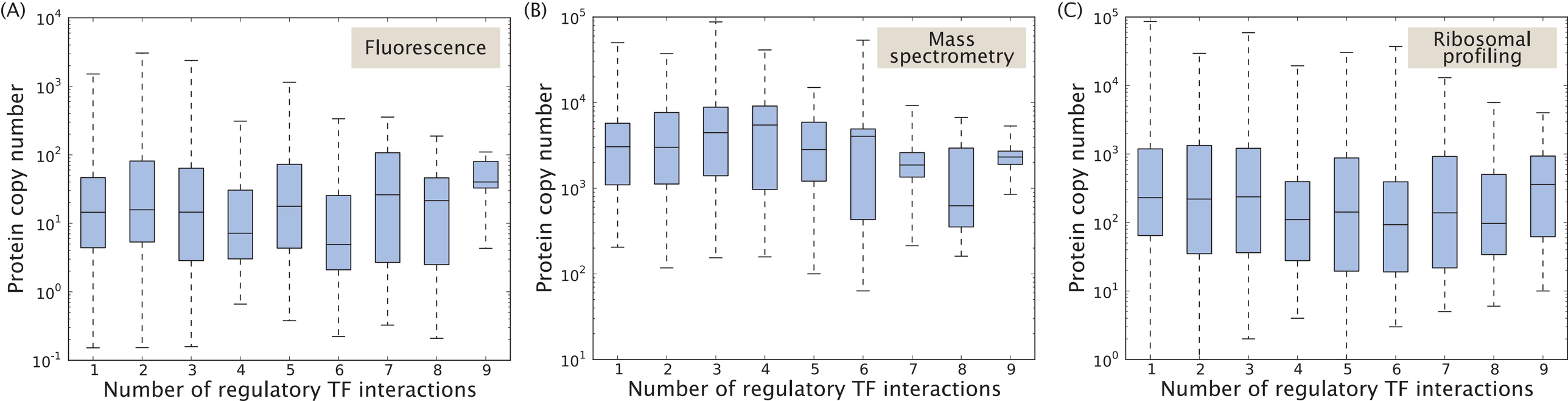

(A)
FlhDC
ycgR
yecR
ygbJK
yhjH
mglBAC
napFDA
GHBC-
ccmABCD
EFGH
ppdAB-
ygdB-
ppdC-
recC
nrfABC
DEFG
mdh
fliE
hydN-
hypF
gltI-
sroC-
gltJKL
fliDST
glpABC
fliAZY
fliLMN
OPQR
fli
FGHIJK
flgBCD
EFGHIJ
flgAMN
flhBAE
(B)
Operons
Genes
Binding sites
Interactions per TF
TFs
10^0
10^1
10^2
10^3
0
50
100
150
200

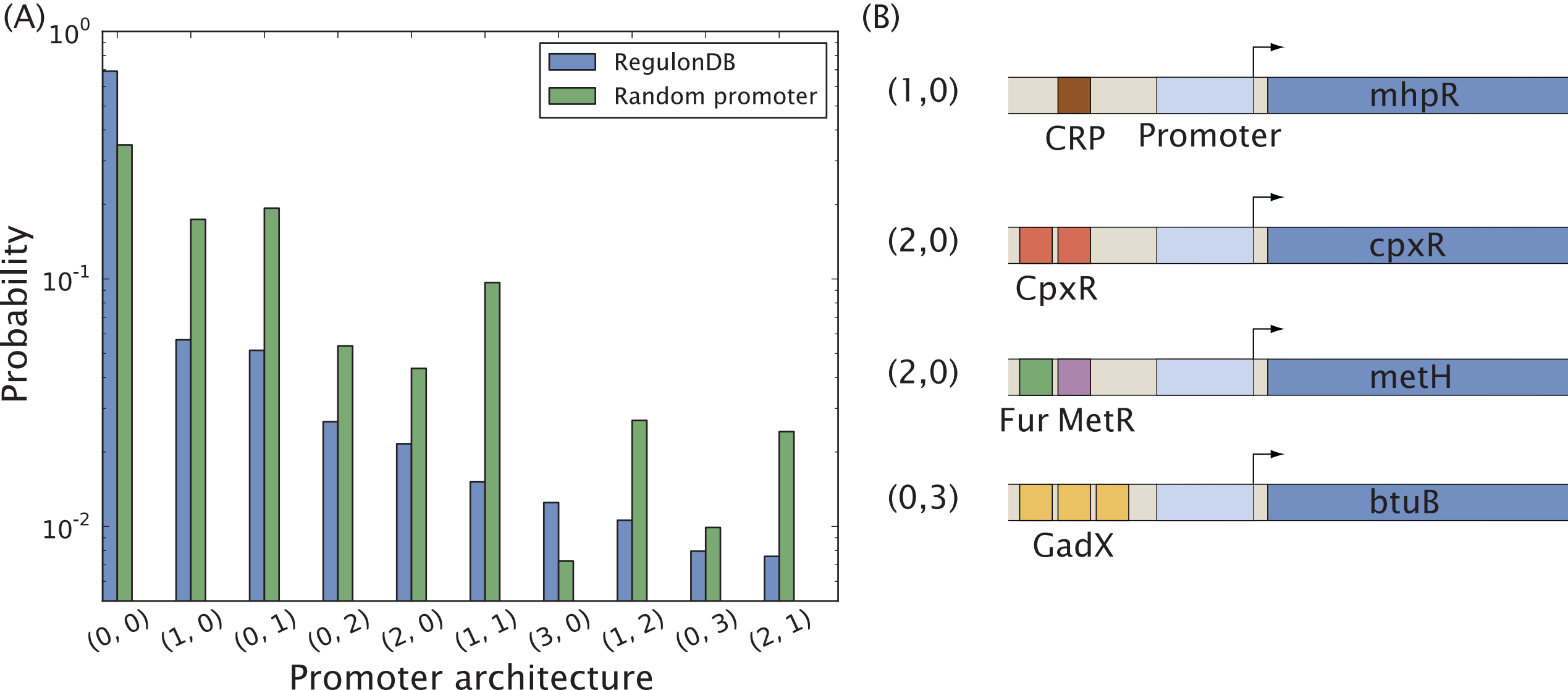

Binding sites
TF types
$590e^{-0.44x}$
$960e^{-0.68x}$
Frequency
10^3
10^2
10^1
0
1
2
3
4
5
6
7
8
Binding sites or TF types per operon

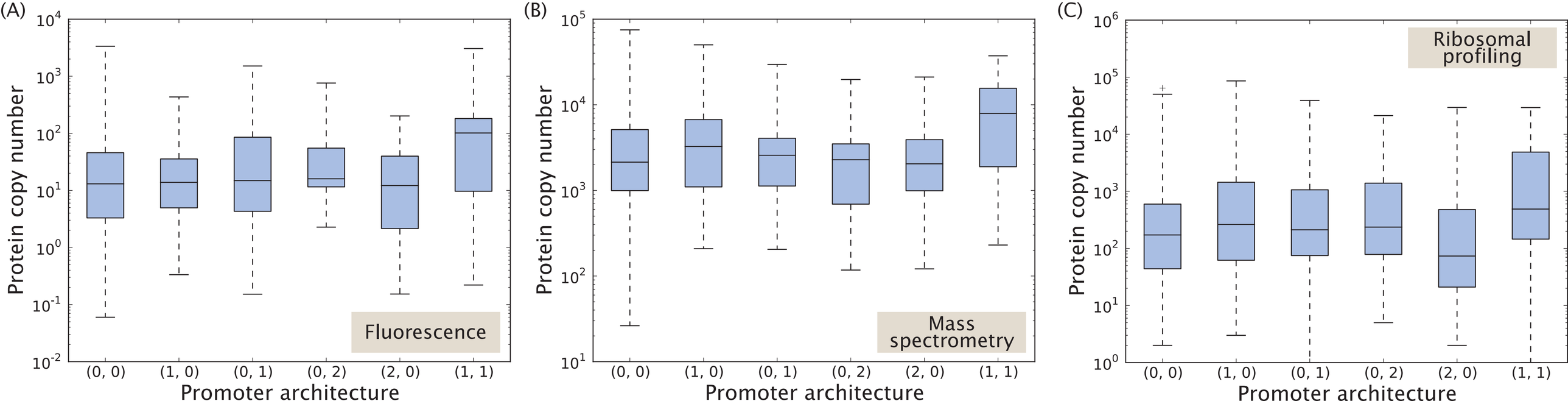

(A)
Probability
Binding sites per operon
RegulonDB
Random promoter
RegulonDB (regulated only)
Random promoter (regulated only)
(B)
GlpR
IHF
Fis
CRP
FNR
glpT
Promoter


OxyR
Fur
fhuF
Promoter

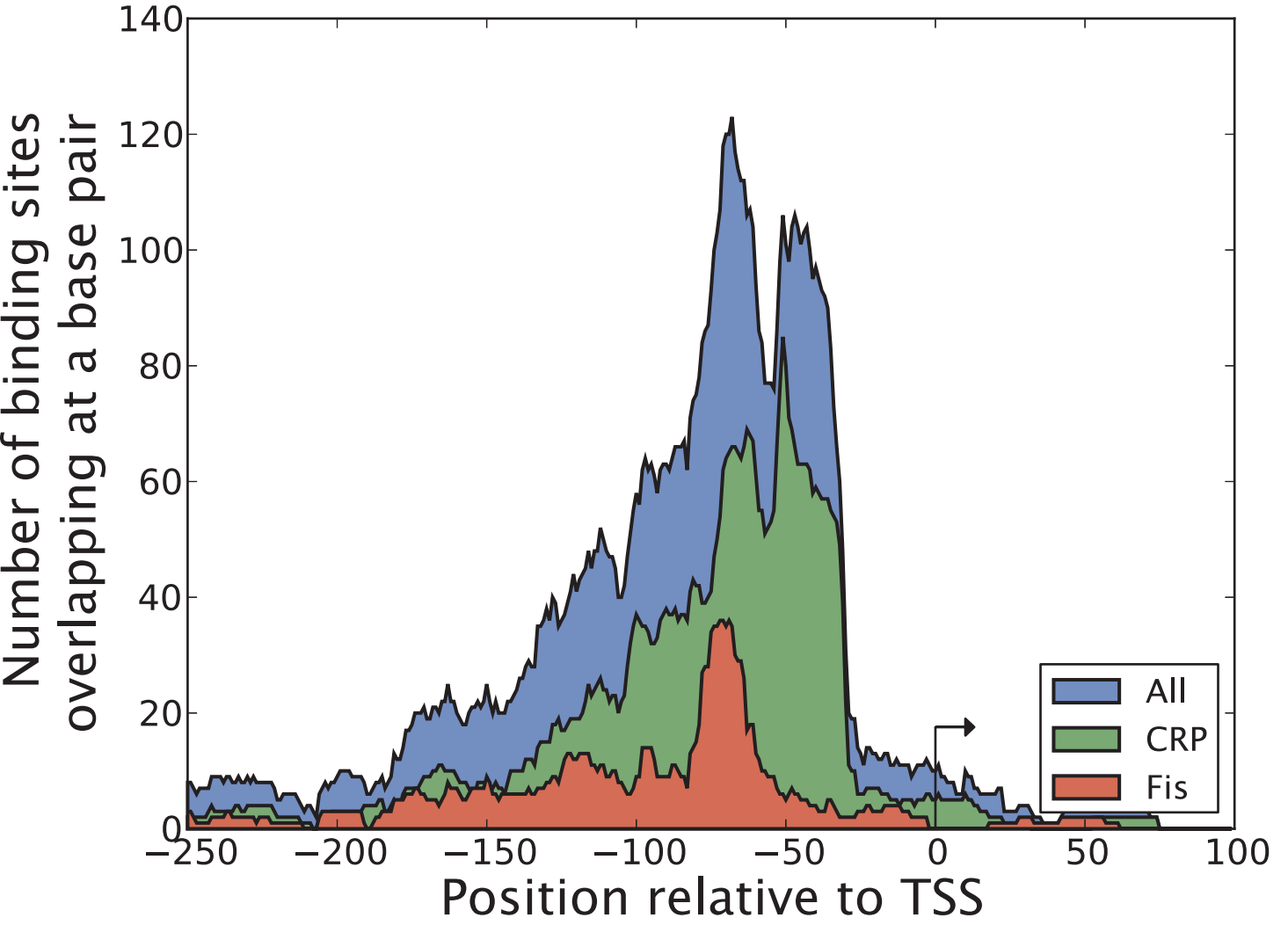

TF copy number
Number of binding sites
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{0}
10^{1}
10^{2}
$y = 2x$
$y = x$
Fluorescence
Mass spectrometry
Ribosomal profiling


(A)
Fluorescence
Protein copy number
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
1 2 3 4 5 6 7 8 9
Number of regulatory TF interactions
(B)
Mass spectrometry
Protein copy number
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
1 2 3 4 5 6 7 8 9
Number of regulatory TF interactions
(C)
Ribosomal profiling
Protein copy number
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
1 2 3 4 5 6 7 8 9
Number of regulatory TF interactions

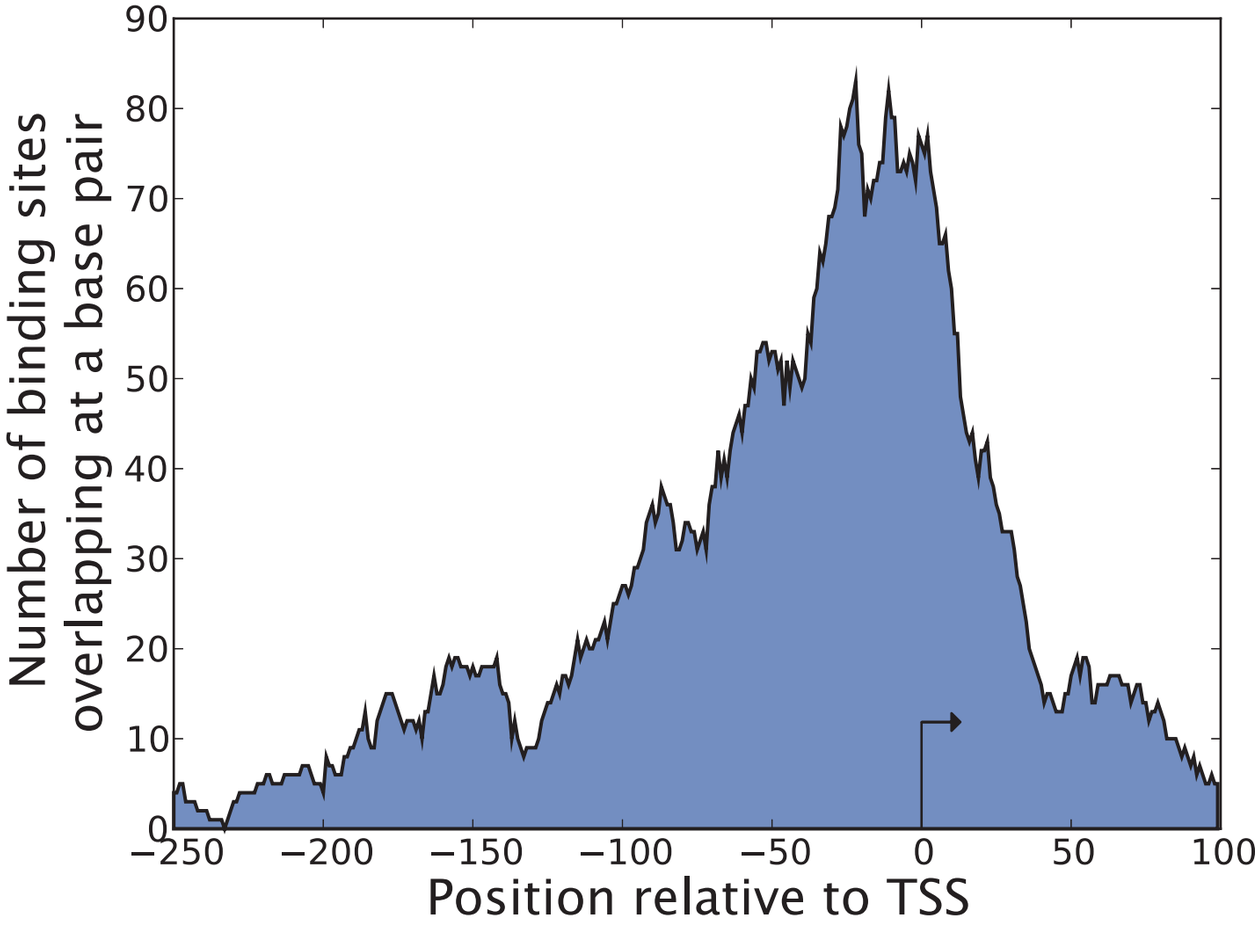

(A)
RegulonDB
Random promoter
Probability
10^{0}
10^{-1}
10^{-2}
(0, 0)
(1, 0)
(0, 1)
(0, 2)
(2, 0)
(1, 1)
(3, 0)
(1, 2)
(0, 3)
(2, 1)
Promoter architecture
(B)
(1,0)
mhpR
CRP
Promoter
(2,0)
cpxR
CpxR
(2,0)
metH
Fur MetR
(0,3)
btuB
GadX


(A)
Protein copy number
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10^{-1}
10^{-2}
Fluorescence
(0, 0)
(1, 0)
(0, 1)
(0, 2)
(2, 0)
(1, 1)
Promoter architecture
(B)
Protein copy number
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
Mass spectrometry
(0, 0)
(1, 0)
(0, 1)
(0, 2)
(2, 0)
(1, 1)
Promoter architecture
(C)
Protein copy number
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
Ribosomal profiling
(0, 0)
(1, 0)
(0, 1)
(0, 2)
(2, 0)
(1, 1)
Promoter architecture

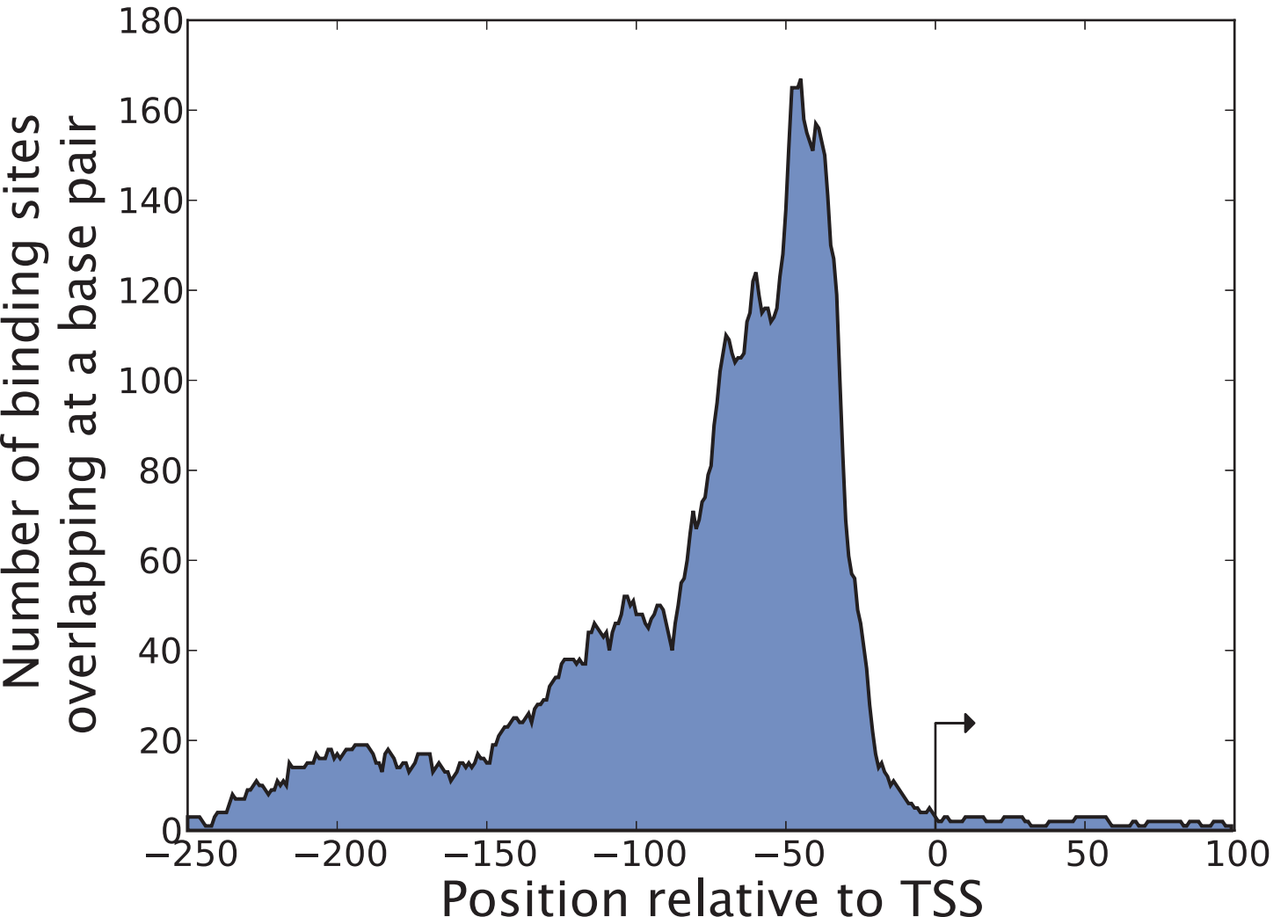
(A) Global activator binding sites



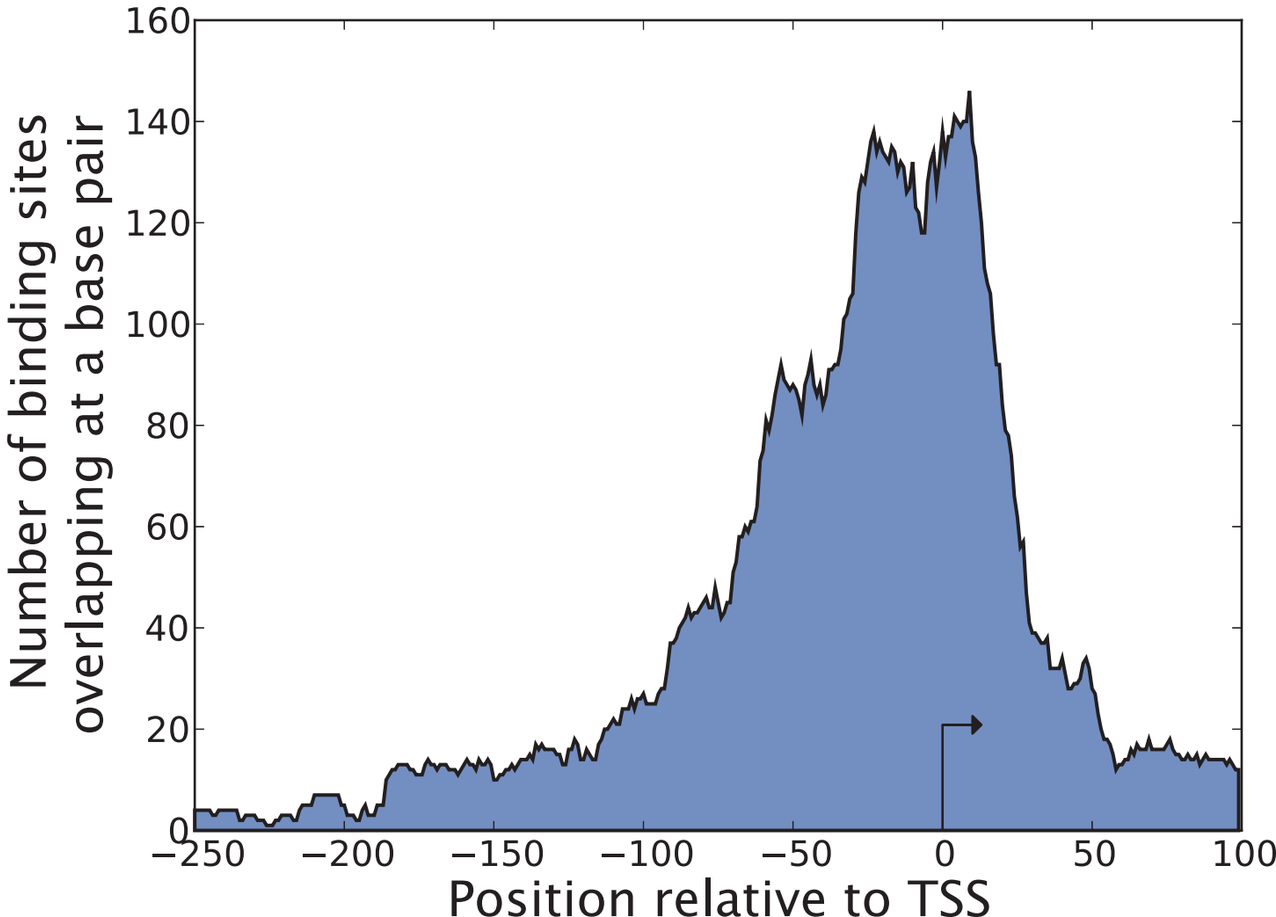
(B) Global repressor binding sites

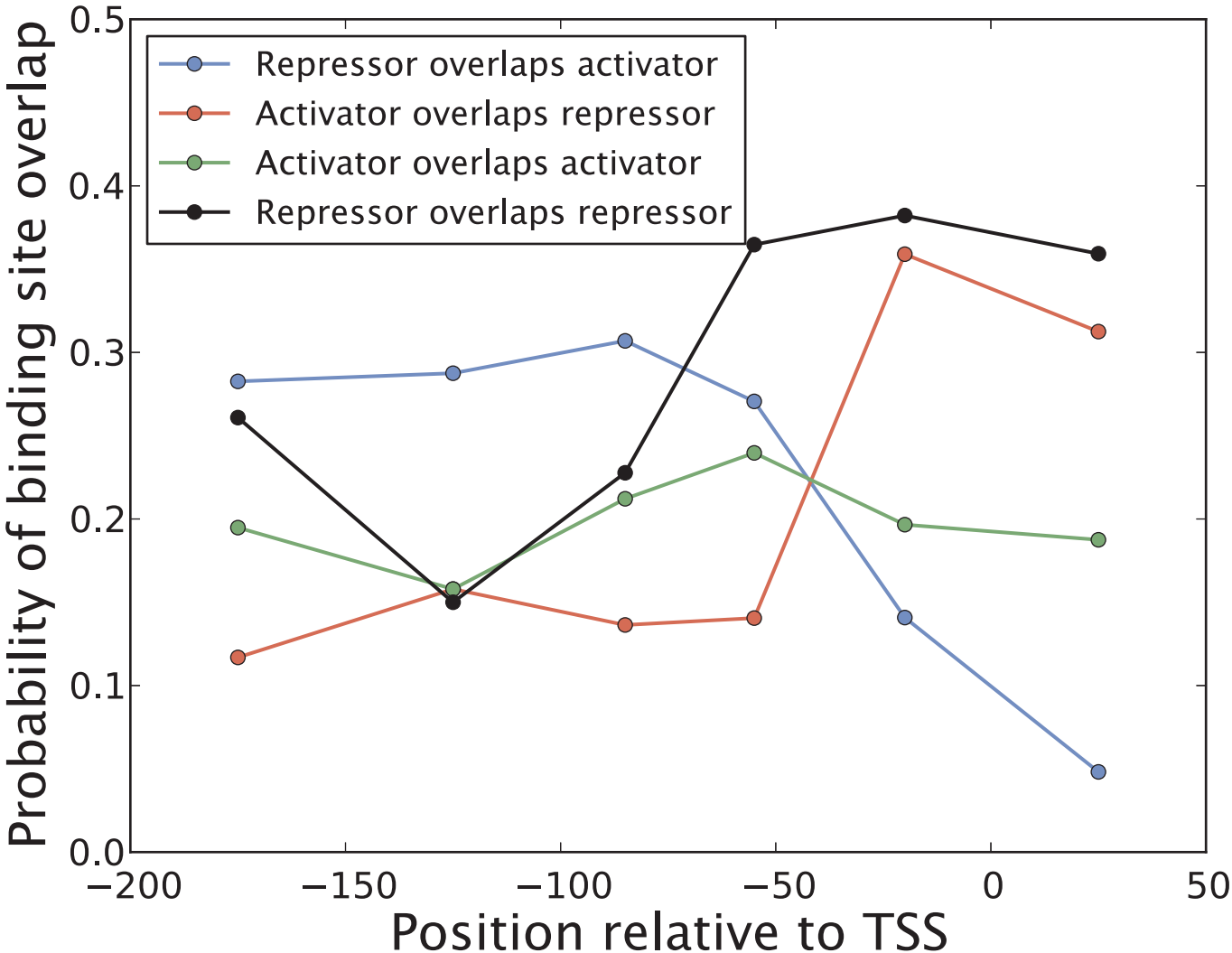


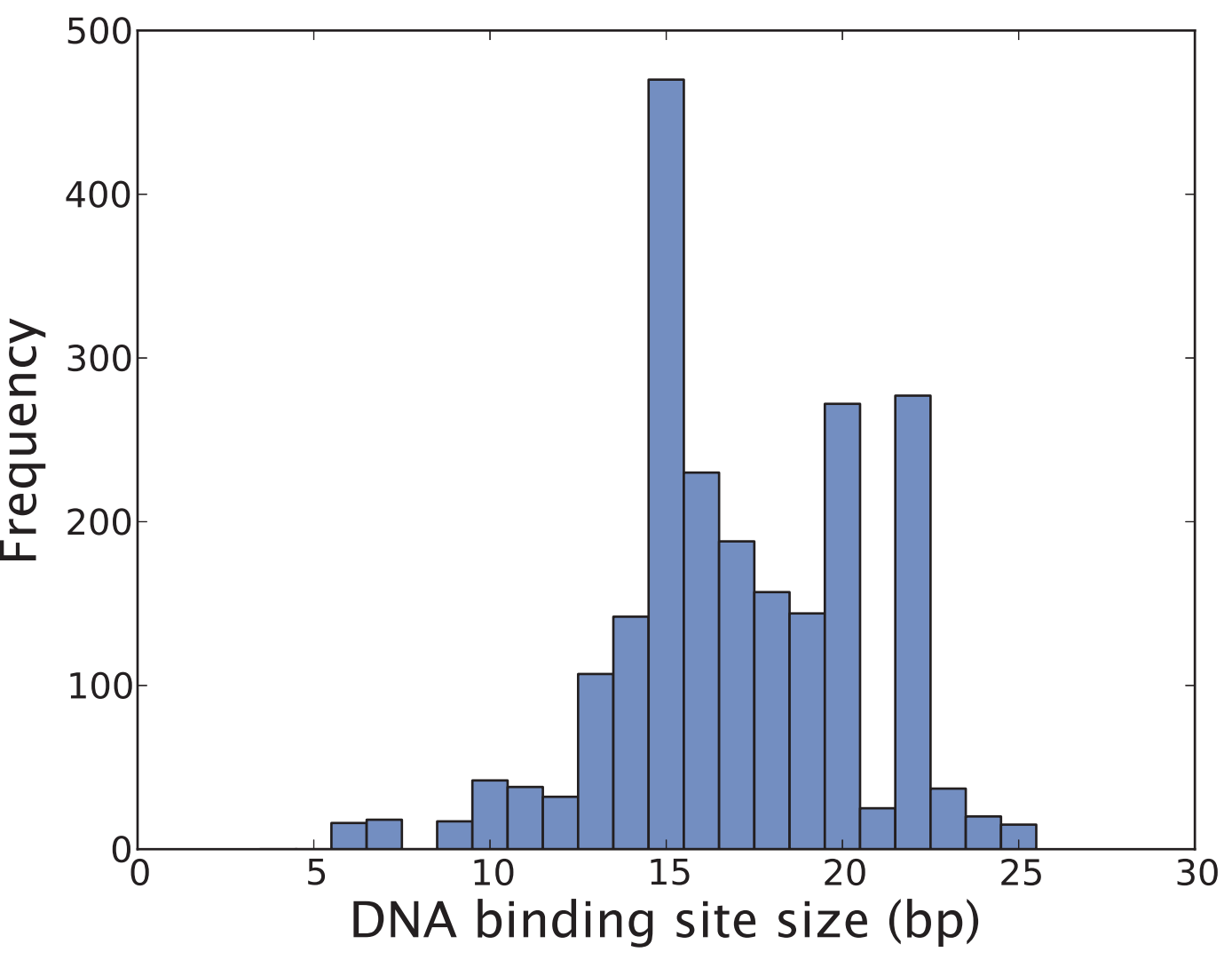
(C) Specific activator binding sites

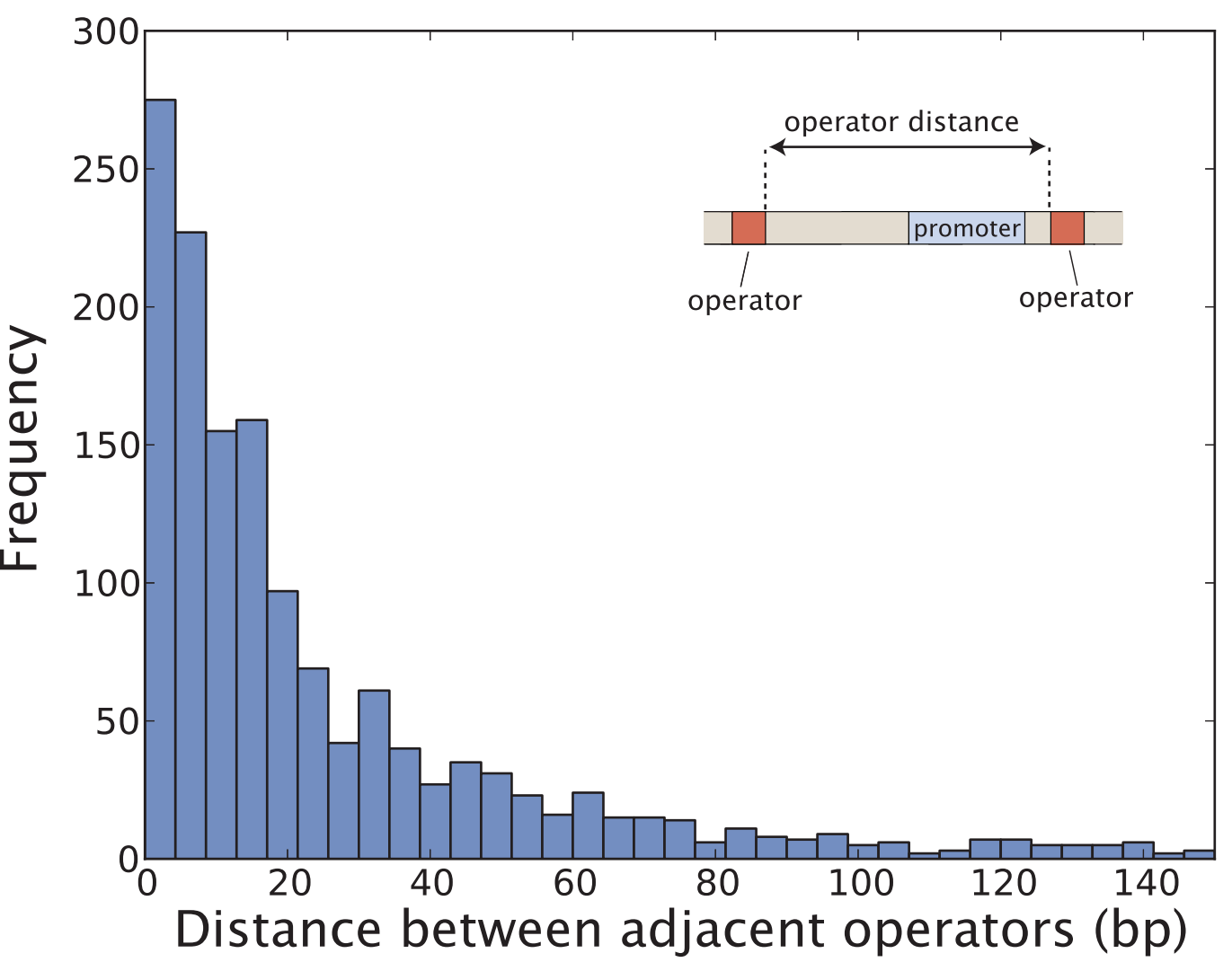


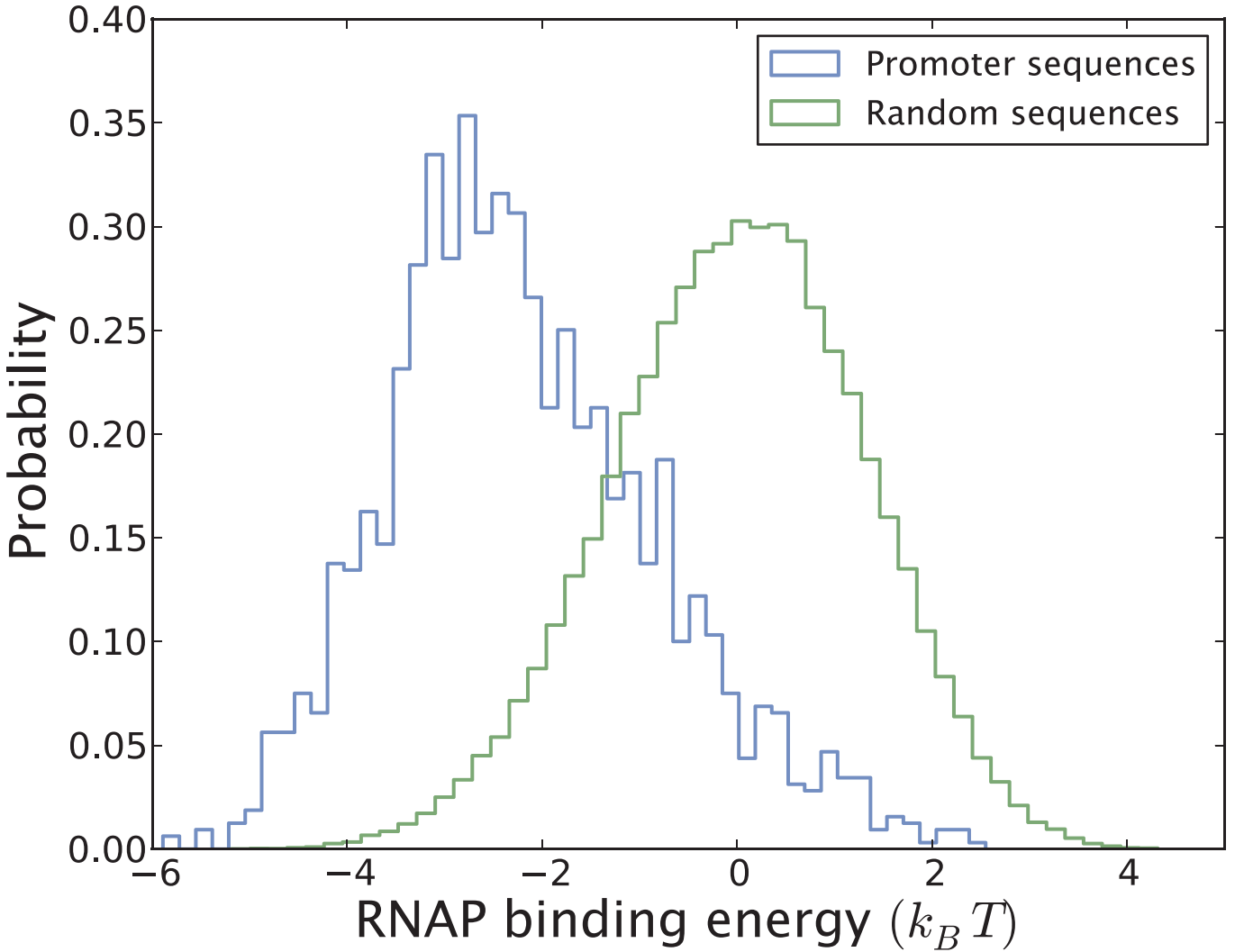
(D) Specific repressor binding sites

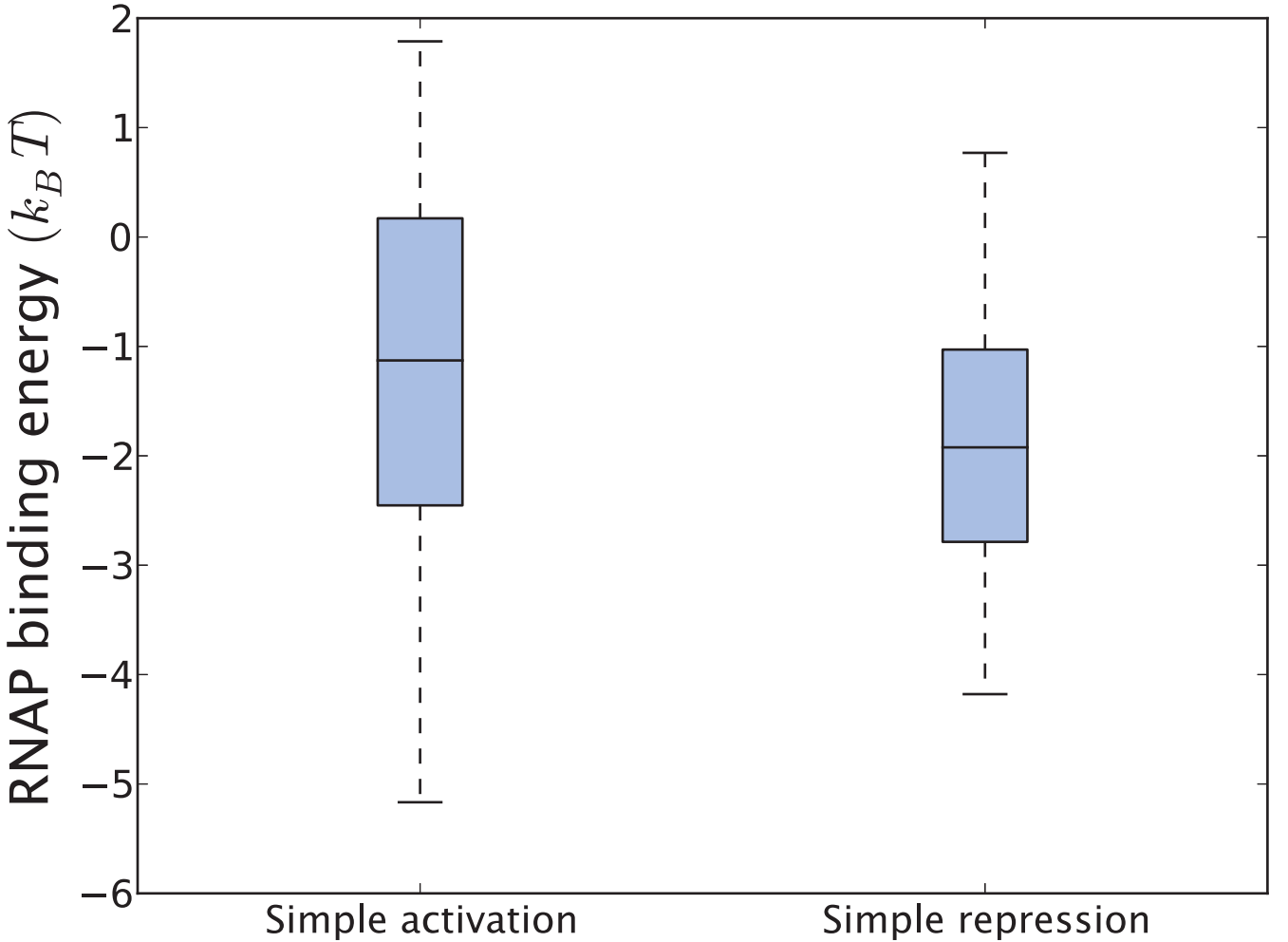


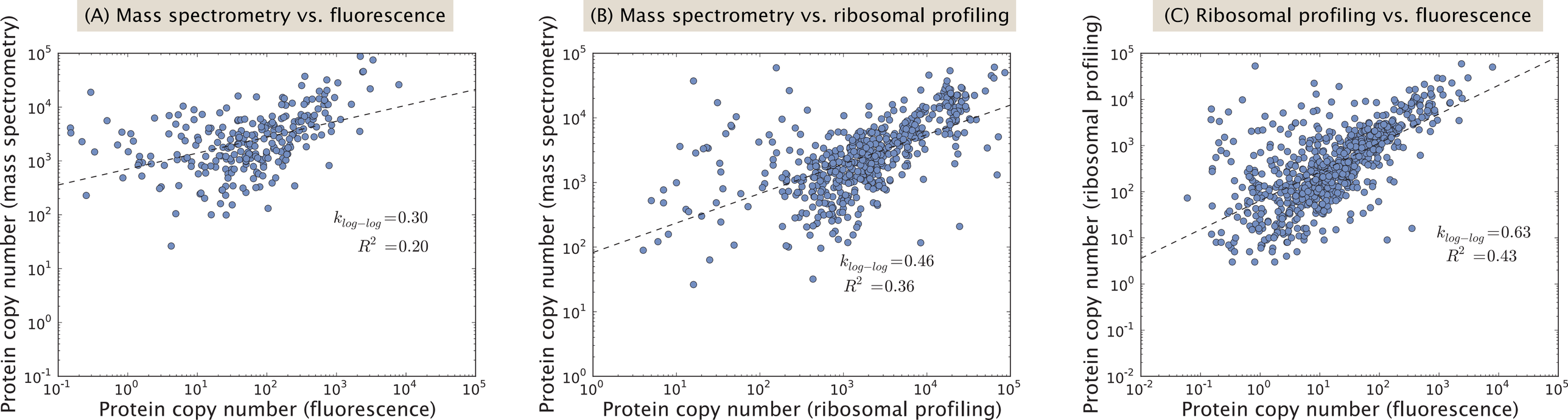

Repressor overlaps activator
Activator overlaps repressor
Activator overlaps activator
Repressor overlaps repressor
Probability of binding site overlap
0.0
0.1
0.2
0.3
0.4
0.5
−200
−150
−100
−50
0
50
Position relative to TSS

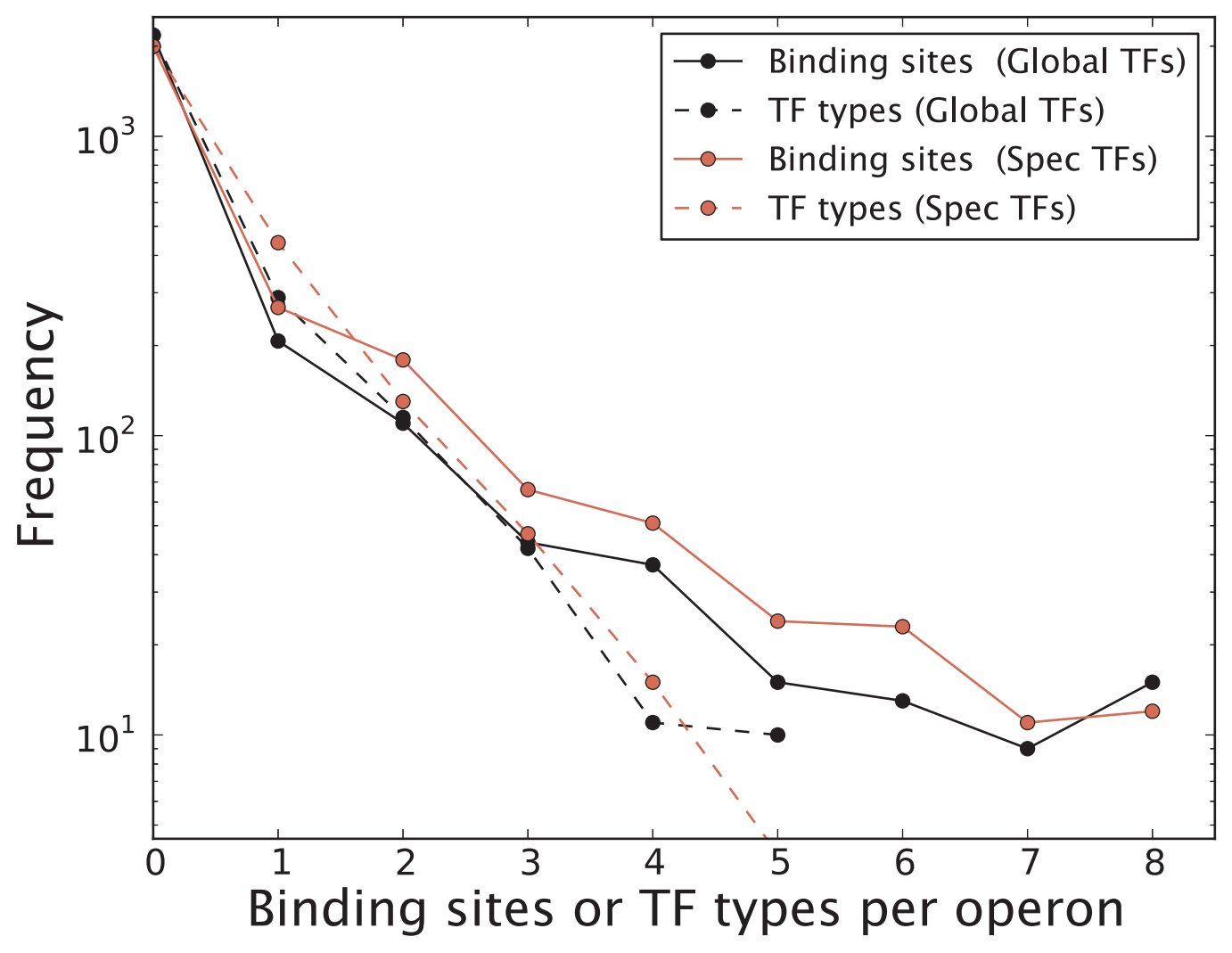

Frequency
500
400
300
200
100
0
0
5
10
15
20
25
30
DNA binding site size (bp)

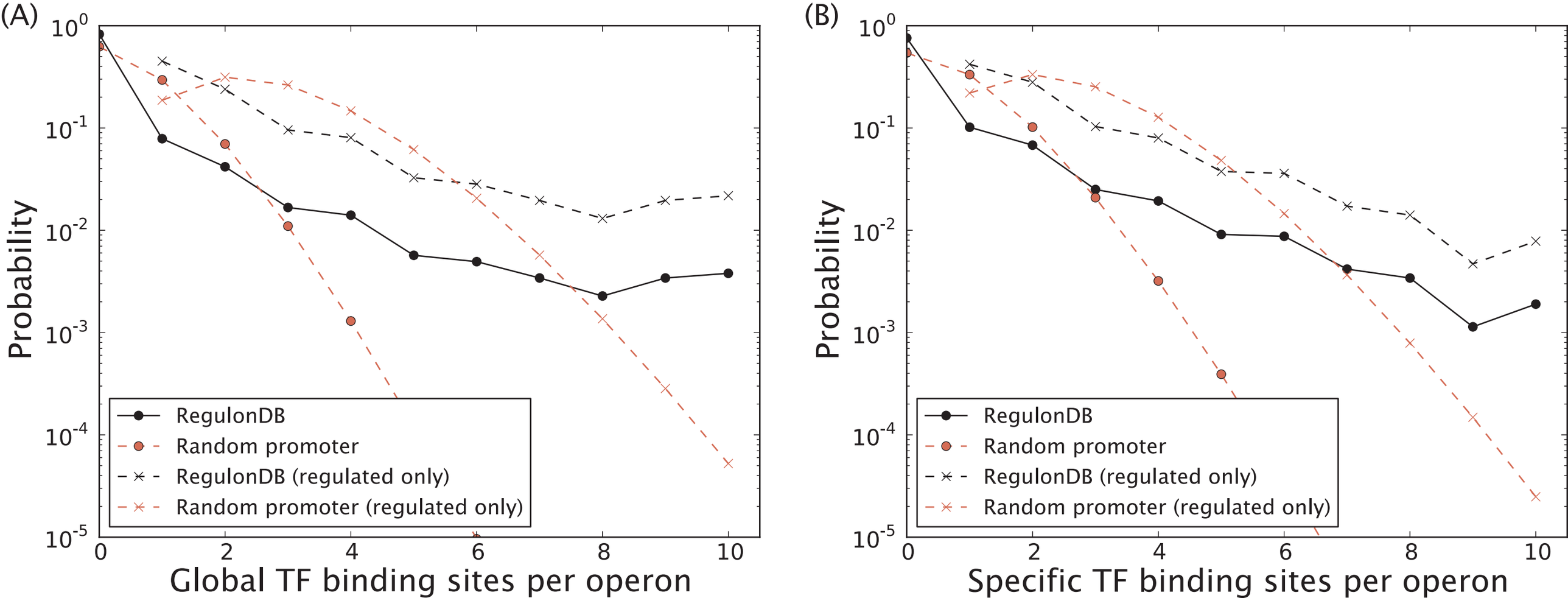

operator distance
promoter
operator
operator
Frequency
300
250
200
150
100
50
0
0
20
40
60
80
100
120
140
Distance between adjacent operators (bp)


Promoter sequences
Random sequences
Probability
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
RNAP binding energy $(k_B T)$


RNAP binding energy ($k_B T$)
2
1
0
−1
−2
−3
−4
−5
−6
Simple activation
Simple repression


(A) Mass spectrometry vs. fluorescence
Protein copy number (mass spectrometry)
Protein copy number (fluorescence)
$k_{log-log}$=0.30
R^2 =0.20
(B) Mass spectrometry vs. ribosomal profiling
Protein copy number (mass spectrometry)
Protein copy number (ribosomal profiling)
$k_{log-log}$=0.46
R^2 =0.36
(C) Ribosomal profiling vs. fluorescence
Protein copy number (ribosomal profiling)
Protein copy number (fluorescence)
$k_{log-log}$=0.63
R^2 =0.43


Binding sites (Global TFs)
TF types (Global TFs)
Binding sites (Spec TFs)
TF types (Spec TFs)
Frequency
10^3
10^2
10^1
0
1
2
3
4
5
6
7
8
Binding sites or TF types per operon


(A)
(B)
Probability
Global TF binding sites per operon
Specific TF binding sites per operon
RegulonDB
Random promoter
RegulonDB (regulated only)
Random promoter (regulated only)