Title
Models to Mechanisms: Leveraging Functional and Evolutionary Information To Describe Regulatory Sequences

Author
Lusk, Richard William

Publication Date
2010

Peer reviewed|Thesis/dissertation

Models To Mechanisms:
Leveraging Functional and Evolutionary Information to Describe Regulatory Sequences

by

Richard William Lusk

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael B. Eisen, Chair
Professor Jasper Rine
Professor Rachel B. Brem
Professor Ian H. Holmes

Fall 2010

Abstract

Models to Mechanisms: Leveraging Functional and Evolutionary Information to Describe Regulatory Sequences

by

Richard William Lusk

Doctor of Philosophy in Molecular & Cell Biology
and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Michael B. Eisen, Chair

The regulation of gene expression is thought to play a critical role in the development of life's complexity and has become one of biology's most intensely-studied areas of research. This study has brought us, in a small set of model systems, a catalog of components and the mechanisms by which these work together to activate and repress expression. A variety of genomic approaches hold the promise of both generalizing these mechanisms and, through the use of statistical models, generating new insights. However, this new wealth of genomic information has provided more raw data than new understanding, due in part to the failure of these statistical models to account for the inherent complexity of biological information.

Here I demonstrate three approaches, spanning analysis of binding sites, promoter regions, and developmental enhancers, to create, gain insight from, and, most importantly, emphasize the need for more biologically informed statistical models.

First, I show how measuring the evolutionary properties of a transcription factor's binding sites can inform the differentiation of those sites from other sequences. That differentiation typically requires the interpretation of a score using a p-value, but, contrary to common usage, I find that the optimal such p-value threshold can differ greatly between transcription factors. Second, I develop a graphical model that can describe and exploit trends in the positioning of transcription factor binding sites within promoters. Binding sites are short and degenerate, not specifying by themselves enough information to mediate the organism's task of promoter recognition. However, I show that these positional trends can greatly increase the information available for recognition, further showing how they can be applied to the bioinformatic promoter recognition problem. Third, I use evolutionary simulations to construct a null model for the relative positioning and conservation of binding sites within developmental enhancers. I use this model to show that much of the evidence supporting the importance of overlapping and clustered sites as functional necessities of enhancer organization can be reproduced as artifacts of constraint on binding site composition alone. Finally, I discuss progress towards testing spatially scrambled enhancers generated from these models in transgenic *Drosophila* embryos.

To Charlie

TABLE OF CONTENTS

Chapter one:

Introduction

One of the most surprising discoveries of the genomic era was the lack of correlation between protein number and organism complexity. Humans, once thought to have 100,000 protein-coding genes or more [1], are now considered to have somewhat less than 25,000. This smaller number is greater than that of the fruit fly, which has around 14,000, but is on the same scale as the microscopic nematode C. elegans, which has 20,000 and is considerably less complex than either. Complexity is now thought to be driven largely by the regulation of these genes [2]. By altering the timing, concentration, and/or location of each protein, selection on changes in gene regulation can explore a vast set of possible phenotypes. Indeed, such changes have been shown to be integral to the diversification and divergence of species [3, 4]. Gene regulation's critical part in the complexity of life, along with its intrinsic roles in development and its pathologies, have made it one of the most intensely studied subjects in biology.

This study, over decades, has produced a great deal of insight on gene regulation's biochemistry and genetics. We know many of the principal components involved in gene expression and have a detailed, if incomplete, understanding of how they work together. We have assembled a large catalog of mechanisms by which genes can be activated or repressed, and in several cases these mechanisms are understood to the limit of molecular detail. For a few model genes, spread across organisms, we have a plausibly comprehensive list of the molecules involved in their regulation and the consequences of each one's disruption. More recently, these insights from small model systems have been applied to the study of gene regulation on a genomic scale.

Genome-wide studies of gene regulation have at least three potential advantages over studies of model systems. First, they address generality: whether principles learned in possibly unusual models truly apply elsewhere. Second, they provide information regarding the great number of genes about which very little is known. Third, and perhaps most important, they hold the potential for gaining new insights that are inherently unavailable to narrower studies: by generating a wealth of data, genome-wide approaches allow problems in gene regulation to be approached statistically. By incorporating our hypotheses about mechanism into statistical models describing these data sets, e.g. how sequence data might translate into expression data, we can approach problems that might be prohibitively difficult to answer satisfactorily with traditional experimental methods.

Yet, thus far, these relatively new approaches have brought more raw data than they have understanding. As discussed in detail below, we now possess a wealth of genomic information stretching from sequence to binding to expression data, and while this information has fueled discovery, the principle question of how sequence dictates expression appears to have only grown more complicated. To some extent, this is a failure of statistics. Statistical models are easiest to apply to data that are independent, unbiased, and unconfounded, and for all of these properties, most biological processes and the means we have to measure them are decidedly not. These shortcomings complicate basic analysis and hinder insight.

Here I discuss my work in three related approaches, spanning analysis of binding sites,

promoter regions, and developmental enhancers, to create, gain insight from, and, most importantly, emphasize the need for more biologically informed statistical models. In this introduction I review our current biochemical and genetic understanding of the function of transcription factors in gene regulation, along with recent computational approaches and how techniques from molecular evolution have been brought to bear on these problems. Then, in the second chapter, I use molecular evolution simulations to frame a discussion of the differences between biological and statistical significance in the context of individual binding sites. In the third chapter, I analyze the differences between promoters bound by different factors. Finally, in the fourth chapter, I show how simple null models of enhancer structure may have led to an incorrect understanding of their functional organization and describe experimental progress towards untangling this question.

Biochemistry and genetics of transcription regulation

Regulatory sequence contains binding sites for proteins, known as transcription factors, that work in combination to dictate the initiation of transcription [5, 6, 7]. In eukaryotes, these sequences can be divided into two classes. The first class defines a group of sequences known as core promoters, which are the nucleation points for transcription initiation [8]. Core promoters contain sites recognized by various subunits of the transcription factor TFIID, which in turn allows the assembly of the remainder of the general transcription machinery: the RNA polymerase, usually the Mediator complex, and a host of other transcription factors required for initiation. However, these core promoters are largely considered to be passive partners in gene regulation [9]. They specify the location of initiation, but some recently discovered exceptions aside [10], they do not appear to play a major role in specifying the location, timing, or abundance of their target gene's transcription.

This role appears to belong to the second class of regulatory sequence. Broadly known across species as cis-regulatory modules, these sequences contain binding sites for transcription factors that, unlike the general transcription factors that bind core promoters, might only be found in particular tissues and only at certain developmental stages or under certain environmental conditions [11]. Cis regulatory elements can thus dictate specific conditions for initiation, and in this manner, can serve as the tools for the creation of complexity and the management of more complex genomes; larger genomes, in general, have more transcription factors per gene than smaller ones [12]. How these sequences integrate the information from their variety of bound transcription factors to activate or repress transcription at the core promoter has become a vibrant and intensely researched question.

Much of this research has focused on the transcription factors binding these sequences and the interactions between them. Individually, these transcription factors are often associated with either activation or repression of transcription, and hence become known as activators or repressors, although others, such as the aptly-named repressor-activator protein (Rap1) in yeast [13], appear to perform both roles in different contexts. These factors appear to affect transcription by recruiting other proteins, coactivators and corepressors, through protein-protein interactions [14, 15]. Some of these cofactors are subunits of TFIID and as such directly link the activator or repressor to the core transcriptional machinery. Others, such as Mediator, form this link indirectly [16]. Finally, a third class uses an altogether different mechanism: altering local chromatin structure to improve or worsen the locus's accessibility. While surely a great diversity

of regulatory mechanisms remain undiscovered, for many factors we have a well developed understanding of how, individually, they act to affect transcription.

Only a very few cis regulatory modules, in their entirety, are understood to a comparable level of detail. Especially in metazoans, these sequences typically bind not one but several transcription factors and each at multiple sites [17]. In turn, each potentially bound factor will vary in concentration according to time and environmental condition. While this combinatorial richness is necessary to specify the great diversity of metazoan body plans and intra-organismal functional operations, its staggering complexity resists case-by-case dissection. In some cases, disruption of any individual site or the creation of small changes in the spacing between sites abolishes regulatory function entirely [18, 19], suggesting that the DNA is serving as a scaffold for the precise assembly of a large multiprotein complex. In others, mutations of activator or repressor sites do not break the enhancer but only constrict or expand, respectively, its activity [20, 21], creating a picture of the enhancer as less a scaffold than an integration point for regulatory information. Finally, some enhancers, such as that for the second stripe of embryonic D. melanogaster even-skipped expression, appear remarkably tolerant of changes to the number and even the identities of the proteins bound to the binding sites they contain [22]. Determining which of these behaviors is most typical requires scaling analysis beyond model genes, which is not feasible for these intensive genetic interrogations. Recent advances in sequencing and other genomic technologies and approaches have made great strides in deciphering enhancer mechanisms, but any discussion of these would be impossible without some treatment of the bioinformatic terms, advances, and challenges inherent to them.

Bioinformatics of regulatory sequence

The most fundamental bioinformatic challenge presented by regulatory sequence is the representation of transcription factor binding sites. With a truly faithful representation, containing all the information available to the factor itself, we should be able to computationally locate binding sites in the genome to the same degree of accuracy that the factor can in the living environment of the cell. Not surprisingly, we continue to fall well short of this mark. However, as we shall see, not only have new biological insights inspired superior statistical binding site representations, but the increasing effectiveness of these tools have also generated valuable hypotheses about regulatory function, from single binding sites to whole cis regulatory modules. In this manner, improvements to our ability to predict binding sites have come hand in hand with a greater biological understanding of how factors recognize their targets and dictate expression patterns.

True binding sites are challenging to represent and discover because they are generally unremarkable features in the genome sequence. A typical eukaryotic binding site might be between six and ten base pairs long [23], yet a six base 'word' in the genome will be found by chance approximately once every few thousand positions, creating, in many cases, more false potential binding sites than there are factors in the cell [24]. Moreover, these sites are degenerate. They are usually not single words but large collections of similar ones that when aligned might only share two or three positions of perfect agreement. This degeneracy makes any kind of representation necessarily involve a degree of complexity.

The earliest representations were intuitive but suffered from severe tradeoffs between sensitivity and specificity [25]. These representations, still in use today, involved making

consensus sequences: lining up all known examples of the binding site and creating a word by stringing together the most common base or bases at each position. While this particular consensus word might not match any individual known site exactly, by allowing a specified number of mismatches, most or all known sites could be recovered. However, as discussed by Stormo [25], the number of mismatches required to match all known binding sites often reduced the specificity of the consensus sequence to the degree that it would identify sites at every few tens of base pairs in the genome.

Clearly, both bioinformaticians and the organism's regulatory system require more recognition capacity to perform their jobs. This capacity could come from two sources: either site-external information, such as co-recruitment by nearby factors, or overlooked information present in the collection of known binding sites itself. Exploring the latter possibility, position weight matrices were developed to capture this overlooked information [26]. These matrices are four rows high, corresponding to the four nucleotides, and as wide as the binding site. They can generate a score for any given potential site by overlaying the site's sequence over the columns of the matrix and summing the row values according to the appropriate nucleotides. The scores are in turn associated with p-values [27, 28], allowing position weight matrices to flexibly translate any given word into a degree of statistical significance.

Known binding sites are used to assign the values of the matrix. There are several methods currently used to assemble these binding sites, ranging from in vitro methods such as SELEX [29], footprinting [30], and protein-binding microarrays [31], to in-vivo methods such as ChIP-chip and ChIP-seq, which will be discussed later in this chapter. The standard method for making this assignment relies on comparing the distribution of bases found at each position in a population of known sites with the distribution of bases found in non-site, 'background' DNA. Briefly put, it assigns each position/nucleotide value in the matrix to be the log probability ratio of drawing that nucleotide from the position-specific binding site distribution or from the background distribution. This method has interesting properties. First, it corrects for skewed background nucleotide frequencies in an intuitive way: GC-rich binding sites will stand out with higher scores in an AT-rich genome. Second, and most useful, this method's scores correspond to maximum probability estimates of the binding energy of the sequence with the factor. Not surprisingly, approaches like these far outperform ones based on consensus sequences.

While these approaches made identification of binding sites much more accurate, they continued to be unreliable. In particular, position weight matrices usually produce a large number of false positive hits: there must still be some information in the system that is unaccounted for. Exploring the former possibility from both experimental and computational perspectives, some groups have researched whether this information could be provided by correlations between positions in the binding site and other nucleotides both inside and outside of the site [32,33,34,35]. These correlations, reflecting the biochemistry of transcription factor structure and the structure of DNA, appear to have an effect at least in some cases, but it appears to be a small one, and the much increased number of training examples needed to train such models precludes their widespread application.

Even if positions within a binding site are not independent, their short length ensures that identical copies will be found throughout the genome, most of them not likely being bound or affecting expression. This implies that accurate identification of binding sites, both from the perspective of the organism and from the perspective of the bioinformatician, must rely on

information outside the binding site sequence alone. Locating this information has become a bioinformatic problem in and of itself: if we are unable to identify functional binding sites precisely, how can we discover the contexts used by the organism to differentiate them? Computationally, this has been approached using two shortcuts: the enrichment and, more recently, conservation of sites and site context. These are discussed at length in the next section.

Genomic approaches to understanding regulatory DNA

A variety of genomic technologies, led by the increased ease of large-scale sequencing, have changed the way that regulatory DNA is studied. These changes are centered on an increased focus on statistical analysis: where pre-genomic analysis generally relied on a few model systems, whole genome sequences revealed tens of thousands of potential regulatory DNAs. This section discusses how advances in genomic technology, with associated advances in bioinformatic and statistical techniques, have both moved forward and introduced new complications to our study of regulatory DNA.

Sequencing

For the study of regulatory DNA, whole genome sequencing had the effect of dramatically increasing, in a relative instant, the number of regulatory elements potentially available for analysis. Complicating this availability was the problem of identification: before studying these regulatory sequences, they needed to be separated from the rest of the sequence. In yeast, this regulatory information is concentrated proximal to the start of the gene, but enhancers in other organisms were already known to function at great distances. From molecular and genetic work, bioinformaticians could leverage several known features of enhancers towards their identification. Foremost among these was the property of binding site density: as most well-studied enhancers contain a large number of transcription factor binding sites, they should appear in the context of non-regulatory sequence as tight clusters of sites. Berman et al [36] identified a set of these clusters and showed that many, when placed in artifical reporter constructs, reproduced the expression patterns of nearby genes. Some enhancers appear to rely on specific spatial constraints between transcription factor binding sites, and these constraints can likewise be used as a tool for enhancer discovery: training a set of these constraints in a set of similarly-expressed genes in D. melanogaster, Erives & Levine [37] were able to locate an enhancer with analogous expression in the distantly-related mosquito genome.

Even as known features of enhancers were used to newly identify regulatory sequences, collections of putative regulatory sequence were being used towards uncovering novel enhancer properties. By integrating over a great number of regulatory sequences across the genome, these approaches carried the promise of discovering the typical functional requirements of enhancers, a direction which had produced seemingly contradictory results in studies of model systems. Using the known affinities of several transcription factors, several groups looked for enriched features in the spacing between their binding sites [38, 39, 40]. They found that binding sites tend to be found in tightly-spaced clusters and, when they are able to, tend to overlap much more often than expected by chance. These results imply that enhancers might rely on tight local interactions between sites, perhaps signifying specific protein-protein interactions, in order to produce specific expression patterns. However, as I discuss extensively in chapter four, the null models used to judge these enrichments appear to be inappropriate.

Genome-wide chromatin immunoprecipitation

To th extent that binding signifies function, the computational problem of enhancer identification can be avoided by the use of whole-genome chromatin immunoprecipitation (ChIP). Relying on a completed genome sequence, ChIP methods pull down DNA bound to a transcription factor of interest and, using either microarrays [41, 42] or, recently, sequencing [43] to determine where in the genome and in what quantity the factor is binding. This technique has two key features. In addition to replacing the unreliable step of computationally locating regulatory regions, it also captures protein-DNA interactions in vivo, which should account for context-dependent effects on binding. A powerful method, in this manner Harbison et al [44] constructed a rough draft of the regulatory map of S. cerevisiae: performing ChIP-chip on most of its transcription factors and locating bound intergenic regions for more than half of them across a number of environmental conditions. Similar work, on a less comprehensive scale, has outlined particular regulatory systems in many other organisms.

This new abundance of binding data has brought with it a new abundance of complications. First among these is the challenge of assigning statistical significance to the binding signal [45]: binding data can be represented as a graph of signal strength overlaid on the genome, and a typical first step of analysis is to choose some threshold over which peaks on this graph will represent statistically significant, 'true', binding events. Yet it is plausible that biologically significant binding information can exist below a statistically significant threshold, complicating and even compromising further analysis. Even once peaks are assigned, resolution is usually insufficient to assign them to a specific binding site.

Finally, it is unclear how binding relates to biological importance: that a region is highly bound does not necessarily imply that it produces a specific expression patten, and even if it does, the extent to which a given gene's expression pattern is functionally important to the organism cannot be made clear by these methods.

Evolution of individual binding sites

ChIP has given us accurate information regarding where factors bind in the genome, but it falls short of assigning biological relevance to these binding events. A bound region might be considered relevant if it dictated a specific expression pattern which increased the fitness of the organism in some way. This fitness advantage is invisible to ChIP: it is unclear what fraction of bound regions drive expression patterns, and our understanding is murkier still of the fraction of expression patterns that are of any functional consequence to the organism. Li et al [46] discovered binding in many regions that were far from their nearest neighboring genes and likely without regulatory activity. Searching for enhancers driving expression in the heart, Blow et al [47] found that a quarter of bound regions did not drive an expression pattern, but searching less specifically, Visel et al [48] found that almost 90% did. However, Berman et al [49] showed that clusters of binding sites, putative enhancers, when artificially placed near a promoter can have expression patterns unrelated to any nearby genes, suggesting that the sites may be bound, produce a pattern, but are nonetheless inconsequent. In order to statistically exploit the vast number of predicted regulatory sequences in the genome, we need methods that, ideally, are able to sort functionally consequent from inconsequent sequences on a large scale.

Methods drawn from molecular evolution are increasingly being used to fill this gap, their

application made possible by the recent abundance of sequence data. By sequencing target loci or whole genomes in closely related species, biologists can observe patterns of nucleotide change between organisms. These patterns should differ depending on the functional consequence of the sequence: simply put, nonfunctional sequence should change relatively quickly, while functionally important sequence should change relatively slowly and perhaps according to predictable patterns.

The evolution of individual binding sites provides an illustrative example. Binding sites typically have positions that vary in their requirements for specific nucleotides: for instance, the first position might require an adenine nucleotide, the second position might usually accept a thiamine but may, with some loss of affinity, accept a cytosine, and the third position might have no nucleotide preference. This property predicts position specific rates in functionally consequent sites. The first position should change slowly, as all mutations will destroy the site's function, and the second and third positions should change at rates that are higher and highest, respectively. Position specific rates were first shown by Moses et al [50], and subsequently shown to increase the accuracy of true binding site prediction. They have subsequently become a part of many binding site determination and prediction programs [51, 52, 53].

Evolution of whole enhancers

Describing the evolution of larger regulatory regions has proven just as useful for differentiating functional sequence. Previously finding success in recovering known enhancers by identifying clusters of binding sites [36], Berman et al extended this analysis by testing the function of many of these predicted enhancer elements in the context of conservation [49]. Those clusters that were also present in a related fly species, D. pseudoobscura, had a much higher fraction that drove a specific expression pattern. Even without knowledge of the affinity of target transcription factors, increased conservation of sequence alone has been shown in some species [54], but not all [55], to be predictive of expression function. With these methods, we can gather examples of regulatory DNAs that are likely to be functionally relevant.

Studying the evolution of enhancers has not only aided their identification, but also generated important hypotheses about how they function. Comparing enhancers from D. melanogaster and several other related species, Ludwig et al [56, 57] noted that many of the binding sites shown to be important in the melanogaster sequence appear to have been destroyed in its orthologs. These disruptions tended to be compensated by the appearance of new binding sites. When these orthologous sequences were placed in the context of a transgenic melanogaster embryo, they drove identical expression patterns as the original. Even very different sequences, both on the primary level and on the level of individual binding sites, could drive identical expression patterns.

This experiment highlighted the property known as binding site turnover [58, 59]. As discussed above, binding sites are short and degenerate, making them liable to appear in neutrally-evolving DNA through random point mutations. These new binding sites have the potential to make old ones functionally redundant. Once redundant, old binding sites are vulnerable to deletion, which creates the effect of a binding site that has moved: a binding site turnover event.

Several groups have studied the patterns of binding site turnover to explore the functional necessities of enhancers: by collecting enhancers that produce the same expression pattern, yet

have different arrangements of binding sites, the importance of particular arrangements can, in principle, be determined. Clusters of binding sites that function together, or binding sites that are critically placed in relation to others, will be harder to make redundant and should therefore be preferentially conserved. Comparing several enhancers across a wide phylogenetic gap, Hare & Peterson et al [60] found, in general, that clustered and overlapping sites were preferentially conserved, supporting the interpretation of the increased enrichment of these features discussed earlier. Similar results were found across the twelve sequenced fly genomes [38]. However, as with the work showing these features' enrichment, this analysis was approached using an inappropriate null model which will be discussed at length in chapter four.

Overview of the approach

Biological discovery is increasingly being driven by genomic technologies and computational methods which necessitate the use of complex statistical models. These models are, to a first approximation, mathematically appropriate, but they are rarely informed by the sophisticated biological processes underlying the systems they are applied to. As I show, this naivete can lead to problems ranging from sub-optimal analysis to the creation of artifactual conclusions. Adapting and correcting these models has thus become a biologist's problem. Here I show progress in informing statistical models with biological information in:

(1) the identification of individual binding sites (chapter two),

(2) the identification of targets of specific transcription factors (chapter three),

(3) a re-interpretation of binding site spacing and conservation data in enhancers (chapter four)

Chapter two:

Use of an evolutionary model to provide evidence for a wide heterogeneity of required affinities between transcription factors and their binding sites in yeast

Abstract

The identification of transcription factor binding sites commonly relies on the interpretation of scores generated by a position weight matrix. These scores are presumed to reflect on the affinity of the transcription factor for the bound sequence. In almost all applications, a cutoff score is chosen to distinguish between functional and non-functional binding sites. This cutoff is generally based on statistical rather than biological criteria. Furthermore, given the variety of transcription factors, it is unlikely that the use of a common statistical threshold for all transcription factors is appropriate. In order to incorporate biological information into the choice of cutoff score, we developed a simple evolutionary model that assumes that transcription factor binding sites evolve to maintain an affinity greater than some factor-specific threshold. We then compared patterns of substitution in binding sites predicted by this model at different thresholds to patterns of substitution observed at sites bound *in vivo* by transcription factors in *S. cerevisiae*. Assuming that the cutoff value that gives the best fit between the observed and predicted values will optimally distinguish functional and non-functional sites, we discovered substantial heterogeneity for appropriate cutoff values among factors. While commonly used thresholds seem appropriate for many factors, some factors appear to function at cutoffs satisfied commonly in the genome. This evidence was corroborated by local patterns of rate variation for examples of stringent and lenient p-value cutoffs. Our analysis further highlights the necessity of taking a factor-specific approach to binding site identification.

Introduction

A gene's expression is governed largely by the differential recruitment of the basal transcription machinery by bound transcription factors [2, 7]. In this way, transcription factor binding sites are fundamental components of the regulatory code, and this code's decipherment is partially a problem of recognizing their location and affinity [61]. These are usually determined using position weight matrices, although a number of more recently developed methods are beginning to become adopted [62]. We use position weight matrices here due to their ease of use with evolutionary analysis and their established theoretical ties with biochemistry. A position weight matrix generates a score comprising the log odds of a given subsequence being drawn from a binding site distribution of nucleotide frequencies vs. an analogous background distribution [25]. The score's p-value is used to determine the location of binding sites: subsequence scores above a predetermined cutoff designate that subsequence to be a binding site, and subsequence scores below the cutoff designate the subsequence to be ignored.

The interpretation of regulatory regions is thus dependent on the choice of the p-value cutoff. However, this choice is not straightforward, although it is commonly made to conform to established but biologically arbitrary statistical standards, e.g. $p < .001$. In addition to assuming that this particular p-value is appropriate, the user here also assumes that a single p-value is appropriate for all transcription factors. Being that score shares an approximately monotonic relationship with affinity [63,64], this implies that the nature of the interaction between different transcription factors and their binding sites is the same. This may not be the case. For example, some transcription factors may require a stronger binding site to compensate for weaker interactions with other transcription machinery, and so a lenient cutoff would be inappropriate. Conversely, the choice of a stringent cutoff could eliminate viable sites of factors that commonly rely on cooperative interactions with other proteins to be recruited to the DNA. A single common standard of significance is a compromise that may not be reasonable.

Ideally, biological information should inform the choice of a p-value and its consequent ramifications in the determination of function. Several recent approaches have well used expression [65] and ChIP-chip [66] data towards understanding binding specificity. Here we take advantage of selective pressure as a third source of information. Tracking selective pressure has the advantage of directly interpreting sequence in terms of its value to the organism in its environment; to a degree, function can be inferred by observing the impact of selection. To this end, we propose a simple selective model of binding site evolution. Selection prevents the fixation of low affinity sites that may not affect expression to a satisfactory level and does not maintain unnecessary high affinity sites. We train the model on the ChIP-chip data available in yeast, and we find evidence for a wide heterogeneity in required binding site affinity between factors. Supporting recent work by Tanay [67], many factors appear to require only weak affinity for function, and we find some evidence that these may rely on cooperative binding to achieve specificity.

Results & Discussion

Definition and training of the affinity-threshold model

In order to use selection as a means to investigate function, a model must be defined to describe how selection acts on functional and non-functional binding site sequence. Our model was created to be the simplest possible for our purposes. We assume that binding sites evolve

independently from other sites in their promoter, but that all sites that bind the same factor evolve equivalently. We interpret a binding site's function in a binary manner: our model supposes that there exists a satisfactory level of expression and that binding site polymorphisms that are able to drive this expression level or greater have equal fitness, while binding site polymorphisms that cannot are deleterious. By assuming that this deleterious effect is large enough to preclude fixation in *S. cerevisiae*, our model imposes an effective threshold on permitted affinity: it does not allow a substitution to occur if it drops the position weight matrix score beneath a given boundary. Analogous reasoning lets us treat repressors identically. By imposing a threshold on permitted affinity and by relying on the assumption that position weight matrix score shares a monotonic relationship with affinity [68], we impose a threshold weight matrix score.

Our purpose in training the model is to find where that threshold lies for each factor, which we accomplish using simulation. For any given threshold and matrix, we simulate the relative rates of substitution that would be expected, and then we compare these rates to empirically determined rates to choose the most appropriate threshold. The simulation is run as follows: we start with the matrix's consensus sequence, and make one mutation according to the neutral HKY [69] model. The sequence's score is evaluated: if it exceeds the threshold, the mutation is considered fixed and the count of substitutions at that position is incremented, and if not, no increment is made and the sequence reverts back to the original sequence. This mutate-select process is repeated. Assuming that the impact of polymorphism is negligible, removing a given fraction of mutations by selection will reduce the substitution rate by that fraction. Thus, the proportion of accepted over total mutations at each position is evaluated to be the rate of mutation relative to the neutral rate.

We use sum-of-squares as a distance metric to compare each affinity-threshold rate distribution to the empirical distribution, and we considered the best-fitting affinity threshold to be the affinity threshold that generates the distribution with the smallest distance to the empirical relative rates.

The affinity-threshold model well describes binding site substitution rates

The Halpern-Bruno model [70] has been incorporated into effective tools for motif discovery [51] and identification, and it has been shown to well describe yeast binding site relative rates of substitution [50]. These rates are also generated by our model, and so we judged our model's accuracy by comparing its performance to the Halpern-Bruno model's performance (fig. 1). We aligned ChIP-chip bound regions and computed summed position-specific rates of substitution for the aggregate binding sites of the 111 transcription factors that met our conservation requirements. We were able to find a threshold at which the affinity-threshold model better resembled the empirical data than the Halpern-Bruno model did for 42 of the 49 factors with adequate training data (see Methods). The affinity-threshold model well approximates the position-specific substitution rates of most factors.

The best-fitting score threshold for a transcription factor's binding sites may correspond to their minimum non-deleterious affinity for that transcription factor. If this minimum is variable and can be found through our evolutionary analysis, then we should be able to detect that variability robustly. To this end, we used a bootstrap to assess the reliability of our predictions, resampling the the aligned sites. Although most transcription factors had large

confidence intervals, they were dispersed over sufficiently wide intervals such that we could form three distinct sets (table 1). We grouped factors with lower bounds greater than 5.9 into a "stringent threshold" set, factors with upper bounds lower than 5.1 into a "lenient threshold" set, and factors with upper bounds lower than 12 and lower bounds greater than -2 into a "medium threshold" set; transcription factors appear to have variable site affinity requirements. We use these sets in all further analysis.

The affinity-threshold model predicts extant score distributions for most factors

If the affinity-threshold model is a reasonable approximation of the evolution of the system, then it should describe other properties of the system beyond the position-specific rate variation of binding sites. One additional prediction of the model is the distribution of binding site scores. For each factor in the groups determined above we sampled the Markov chain and computed the mean binding site score under the affinity-threshold model. We compared this to the average maximum score for that transcription factor in ChIP-chip bound regions (fig. 2). Although it had a downward bias, the affinity-threshold model predicted the extant distribution of stringent- and medium- threshold transcription factor binding sites. However, it fared worse with the lenient-threshold binding sites, suggesting that the evolution of these sites may not operate within the simplifying bounds of the model, i.e. perhaps their evolution is governed by a more complex fitness landscape instead of our stepwise plateau. Nevertheless, average maximum scores in bound regions for these factors are still found commonly in the genome.

Stringent- and lenient-threshold binding sites have distinct pattens of local evolution

The lenient set of transcription factors allows for binding sites that would be found often by chance in the genome. If this lenient affinity is truly sufficient, these transcription factors may rely on other bound proteins to separate desired from undesired binding sites. In contrast, sites meeting the affinity threshold for stringent-threshold transcription factors should be high-occupancy sites without a need for additional information due to their strong predicted affinity.

To investigate this hypothesis, we counted the average number of different transcription factors bound at each promoter for each of the factors used in the Harbison et al ChIP-chip experiments. Let ``lenient-group sites'' refer to sites bound by lenient-threshold transcription factors (e.g. Sut1p, table 1), and let ``medium-group'' and ``stringent group'' sites be defined similarly. As expected, the stringent and lenient groups were separated, the lenient group promoters having just under three more unique bound factors per promoter for each of three binding significance cutoffs. However, the medium and lenient groups were not well separated.

We used the variation in local substitution patterns to determine whether medium and lenient group factors could be distinguished by an enrichment of local binding events. While medium and lenient group sites have similar numbers of different transcription factors bound to promoters that they also bind, lenient group sites will have a higher density of other binding sites immediately surrounding theirs if recruitment by other proteins is necessary for their function. This density should be reflected in the local pattern of evolution, as the sequence will be comparatively restrained.

We calculated rates of substitution surrounding the binding sites of stringent-, medium-, and lenient-threshold transcription factors. All transcription factors in each set were pooled and the rate of substitution was calculated and summed by distance to the transcription factor edge.

All three sets have a reduced rate of substitution at the position adjacent to the binding site (fig. 3a), suggesting that some of these weight matrices do not describe the entire factor. Lenient group sites have a depressed rate of substitution relative to the areas surrounding the medium and stringent group sites (fig. 3b, $p < 2e\text{-}16$,chi**2= 160.8, 1df), consistent with a hypothesis of increased local binding. In contrast, the regions surrounding stringent group sites are marked by a shoulder of increased substitution rate (fig. 3a). This shoulder suggests a model in which high-affinity sites sterically inhibit transcription factors from binding to adjacent regions, preventing them from being used as regulatory material. The stringent and lenient group sites are distinguished by their expected patterns of local substitution rate variation.

Transcription factors may best interact if they are on the same side of the DNA [71,72,73], suggesting that binding sites of interacting factors should be phased at approximately 10.4 base pairs to match the periodicity of the double helix, although this will vary according to the particular nature of interaction between the two proteins. If binding sites coordinated in this manner, the substitution rate should match this periodicity. We evaluated the fit of a model that allowed for a 10.4 base pair periodicity in the rate, although the noted variability between interacting factors will reduce the quality of this match. We fit the twenty base region ten bases from the edge of the transcription factor, allowing for two turns of the DNA while avoiding possible occluding effects of the original bound factor. The regions local to lenient group sites fit this model significantly better than they fit a uniform rate model (fig. 3c, p=.0053, chi**2=10.53, 2df), while the regions surrounding medium and stringent group sites did not.

Conclusion

We developed a simple model of binding site evolution to investigate the possibility of differences in transcriptions factors' requirements for binding site affinity. Unlike other models of binding site evolution, the affinity-threshold model is geared toward understanding the transcription factor itself rather than its binding sites. The model was used to create three groups of transcription factors with stringent, lenient, and intermediate requirements for binding site affinity, and these groups were supported by the extant distribution of binding sites and their distinctive patterns of localized substitution rate. We note that some factors appear to evolve and exist at thresholds that poorly distinguish their binding sites from background sequence, perhaps making consideration of context essential for their accurate identification.

Methods

Rate of binding site evolution

We downloaded the *S. cerevisiae* sequences used in the Harbison et al [44] study and used bi-directional best FASTA [74] hits ($p < 1e\text{-}5$) to find the orthologous subsequences in *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus* contigs available at SGD [75]. We aligned the sequences using Mlagan [76].

We obtained ChIP-chip binding data from Harbison et al, using all available conditions for each factor. We used a binding p-value cutoff of .001 to determine binding, but the analysis was fairly robust to using different cutoffs: we also calculated rates of evolution for of transcription factor binding sites for binding p-values of .005 and .0001 and observed similar groups, although some stringent-threshold factors were lowered to the medium-threshold group using the former data set. We downloaded weight matrices for 124 factors [66], and we used

Patser [77] to designate the highest-scoring subsequence(s) within each bound locus to be the subsequence responsible for binding. This choice precludes the inclusion of many functional weak sites, but we wished to minimize the impact of non-functional sites. Alignment errors, binding site turnover, and changes in cis-regulation all will introduce neutral sequence evolution into the model training data, biasing our choice of threshold downward. In particular, Borneman et al [78] highlighted rapid changes in binding for two transcription factors across three yeast species. We hoped to minimize the impact of such by imposing minimal criteria for conservation: we discarded alignments with gaps and alignments containing a sequence with a score beneath zero. We used maximum parsimony for all determinations of substitution rate. Although progress has been made towards determining the neutral mutation processes in *S. cerevisiae* intergenic sequence [79], we wished to avoid remaining uncertainties and so in all cases we compared relative rates within the binding site instead of absolute rates. We did not further analyze transcription factors for which we were unable to train on at least two mutations per position. We calculated the Halpern-Bruno rates according to the method described in Moses et al [50].

Simulation of the affinity-threshold model

We simulated the affinity-threshold model for a wide range of thresholds for each of the 124 weight matrices described by MacIsaac et al. We calculated position-specific substitution rates for score thresholds between -10 and the position weight matrix's maximum in increments of 0.1. This process starts with the consensus sequence and is run for eighteen million iterations. We determined \$95\%\$ bootstrap confidence intervals of the best-fitting threshold by finding the best-fitting affinity threshold for each of 10,000 resamples of the aligned binding sites. Software will be available from http://rana.lbl.gov/\$\sim\$rlusk/PSB2008/.

Predicted equilibrium distribution of scores

We sampled every 20,000th sequence generated by the Markov chain for the best-fitting affinity threshold model for each transcription factor in the three groups. We compared the mean score of these sequences with the mean maximum score of the sequences meeting a \$p<.001\$ ChIP-chip binding cutoff.

Periodicity testing

We evaluated two nested models against the 10-30 base pair region surrounding each binding site. The first supposed a uniform rate alpha across the region to determine k_p Poisson-distributed mutation events at each position p, and the second added a periodicity of 10.4 to this rate with magnitude beta and phase gamma. t_p is the number of gapless alignment columns at that position. The maximum likelihood parameters were discovered by direct search.

$$L(k|\alpha,\beta,\gamma;t)=\prod_{p=10}^{30}\frac{e^{-f(\alpha,\beta,\gamma)t_p}f(\alpha,\beta,\gamma)^{k_p}}{k_p!}$$

$$f(\alpha, \beta, \gamma) = \left\{1 + \beta \sin\left(2\pi \frac{p - \gamma}{10.4}\right)\right\} \alpha$$

Significance was determined using a likelihood ratio test with beta either allowed to fluctuate between zero and one or held to zero.

Figure 1. Position specific rate variation and model predictions for (a) Fkh2, (b) Fhl1, and (c) Aft2: relative rate vs. position in site. The black line marks the empirical rates, the dashed line marks the Halpern-Bruno predicted rates, and grey line marks the best-fitting affinity-threshold. The grey bar contains the set of rates predicted by all affinity thresholds within the factor's 95% confidence interval.

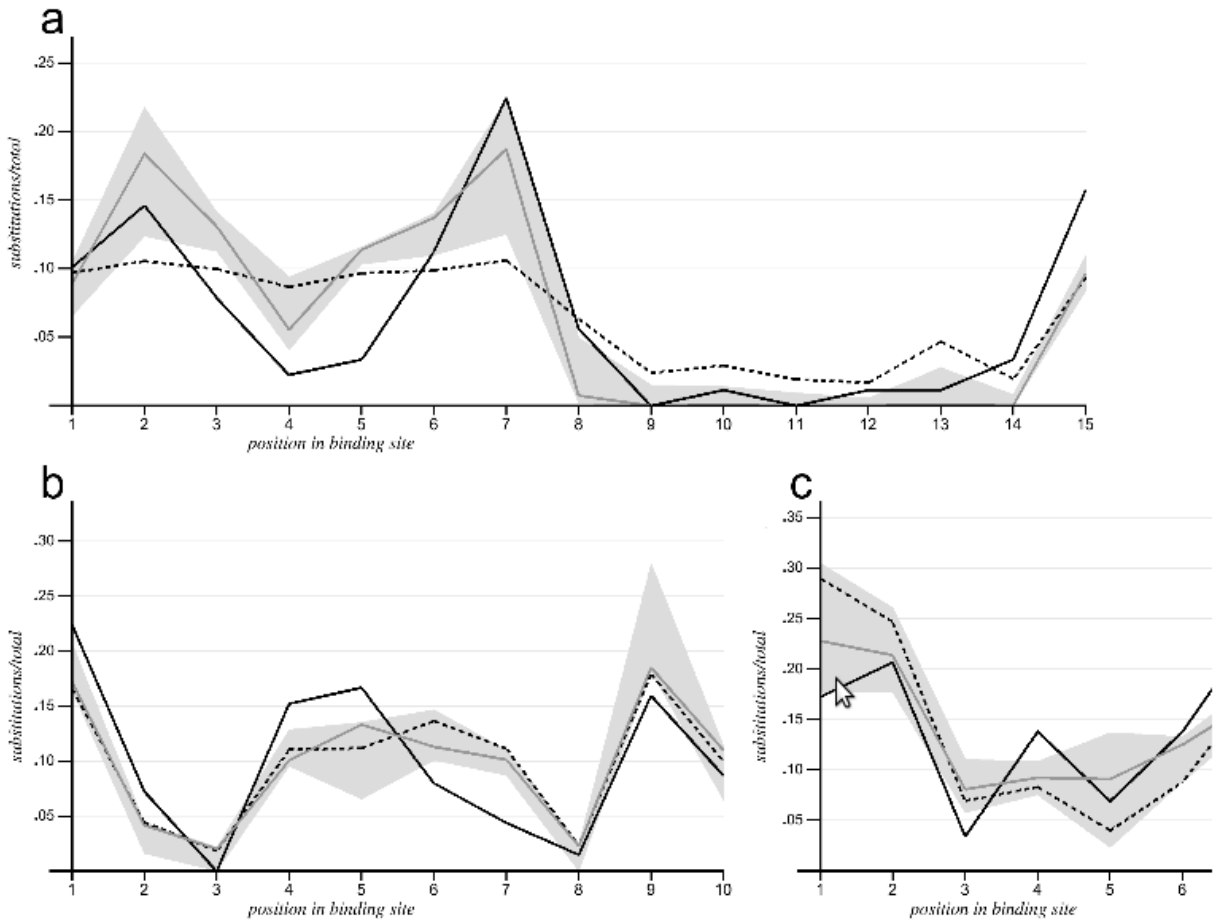


Figure 2. Predicted average score at best-fitting affinity threshold vs. average maximum score in ChIP-chip bound regions. Stringent-, medium-, and lenient-threshold transcription factors presented as black, dark grey, and light grey dots, respectively.

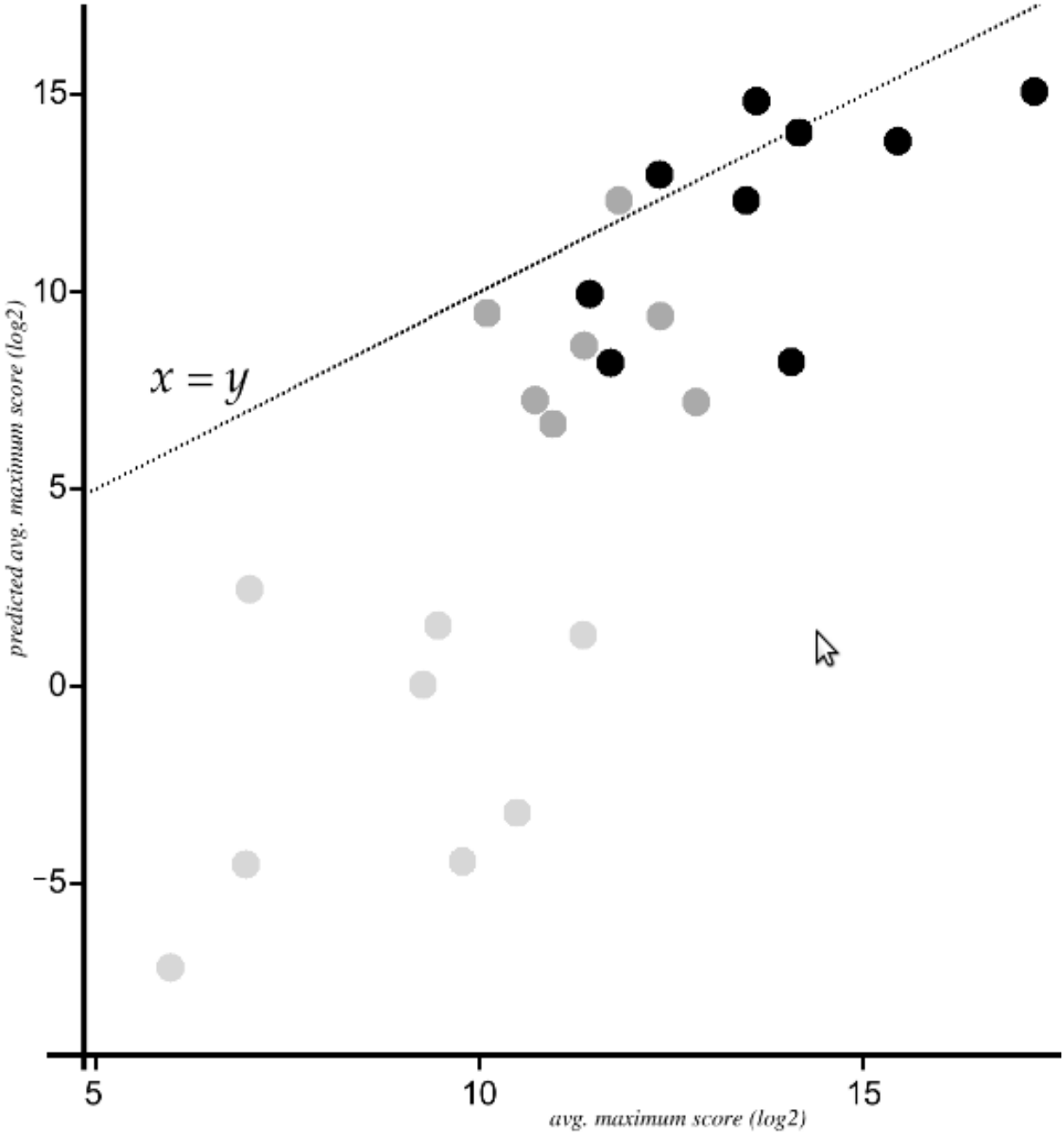


Figure 3. Local rate of substitution (*subst/site*) vs distance to binding site edge (bp).
The solid, dotted, and dot-dashed lines mark the local rates surrounding stringent-, medium-, and lenient-affinity group transcription factor binding sites. In (c), the grey line marks the predicted periodic rate of evolution near lenient-affinity group sites.

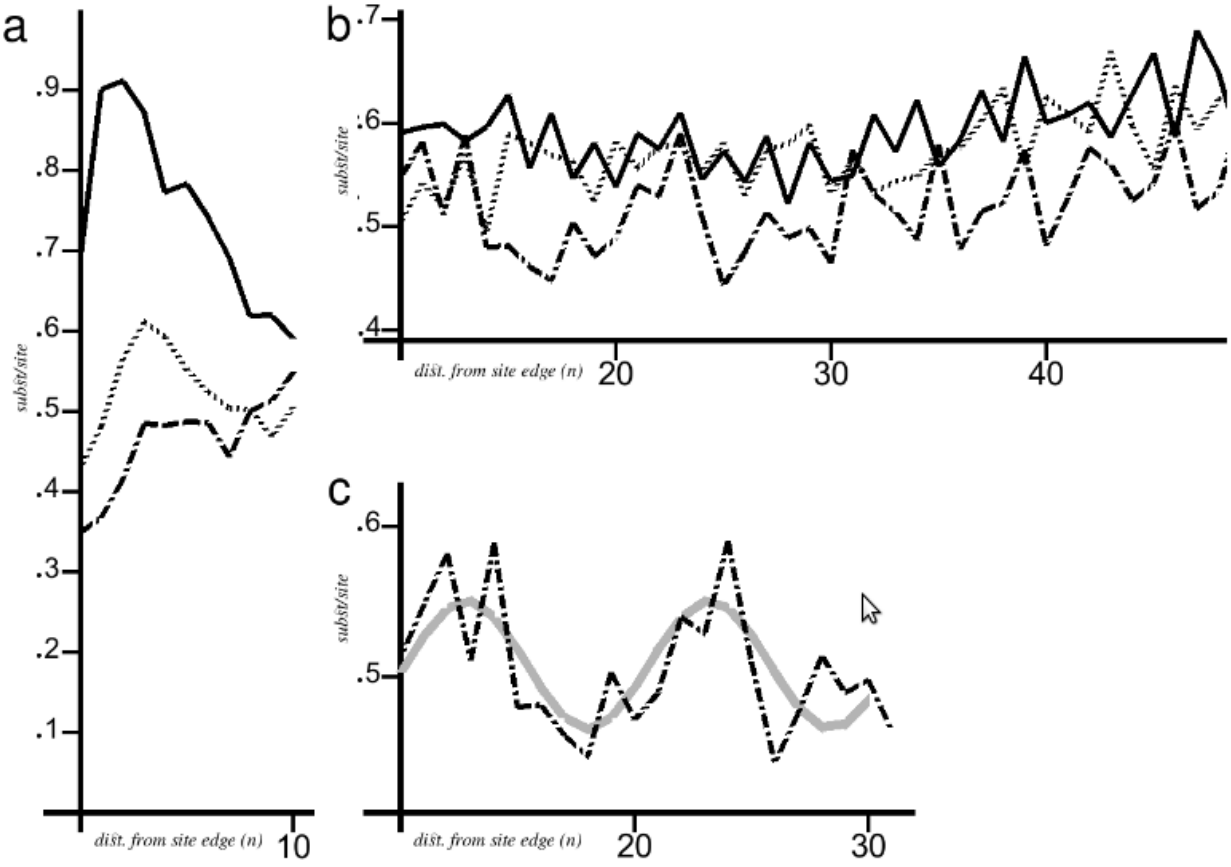


Table 1. Affinity threshold confidence intervals and corresponding site prevalence for transcription factors in the stringent (left), medium (middle), and lenient (right) threshold groups.

	CI^{a}	$Prev.^{b}$		CI^{a}	$Prev.^{b}$		CI^{a}	$Prev.^{b}$
Reb1p	8.3-11.1	.226-.117	Cin5p	−0.4-8.5	.997-.294	Sut1p	−9.9-4.2	.988-.845
Bas1p	5.8-13.6	.566-.005	Mbp1p	2.7-11.7	.793-.059	Aft2p	−9.8-4.2	.988-.794
Fkh2p	8.1-15.2	.497-.003	Fhl1p	4.2-11.3	.702-.048	Phd1p	−9.8-5.1	.998-.867
Cbf1p	6.2-12.0	.219-.028	Gcn4p	4.0-10.6	.682-.080	Ace2p	−9.9-−0.8	.999-.999
Abf1p	11.0-12.9	.108-.075	Swi6p	3.8-9.9	.854-.166	Yap6p	−9.9-4.2	.993-.909
Sum1p	6.2-14.5	.484-.009	Ste12p	1.0-6.5	.997-.705	Adr1p	−9.5-2.3	.991-.856
Tye7p	8.6-11.3	.183-.037	Nrg1p	−1.3-7.0	.968-.388	Hap5p	−9.4-−2.1	.993-.993
Mcm1p	8.7-19.5	.133-.002				Mot3p	−2.9-5.1	.996-.595
Hap4p	11.0-14.9	.059-.003						

Note: [a] 95% confidence interval, log base two scores

[b] Prevalence: first and second quantities are the fraction of all promoters containing a site meeting the lower and upper bounds of the CI, respectively

Table 2. Average number of binding sites per promoter, grouped by best-fit affinity threshold and ChIP-chip binding p-value.

Group	$p < X$		
	.005	.001	.0001
Stringent	7.78	4.74	3.33
Medium	10.30	7.09	5.13
Lenient	10.73	7.59	6.25

Chapter three:

Spatial promoter recognition signatures enhance transcription factor specificity in yeast

Abstract

Transcription factors recognize their target promoters through their sequence-specific binding affinity. However, due to their short length and high degeneracy, transcription factor binding sites in yeast have been shown in theory and in practice to be insufficient to mediate this recognition alone. Some factors require their binding sites to exist in a particular arrangement or context in order to function, and recent computational work has shown that these patterns, which we term promoter recognition signatures, may be a common property of yeast transcription factors. Here we train spatial models of binding site positioning, uncovering this spatial information for a large fraction of transcription factors. Unlike previous work, we focus on the characteristics of promoters rather than individual binding sites, allowing us to uncover transcription-factor-characteristic differences in site density, which appears to play an important role in specificity. We show that these signatures allow transcription factors with substantial differences in binding site specificity to share similar promoter specificities. We illustrate how these signatures greatly increase the information available to the organism for promoter recognition. Finally, we show how these spatial signatures can be brought to bear upon the bioinformatic problem of target differentiation. Signature-derived scores show superior performance than those derived from models that do not take into account spatial information and, in an appreciable fraction of cases, they outperform ChIP-chip binding predictions.

Introduction

The regulation of gene expression is mediated by proteins called transcription factors, which bind specific gene promoters and work in conjunction with other proteins to either activate or repress expression [5,6,7]. How these factors differentiate their proper targets from the rest of the genome has become a vibrant question in the study of gene expression.

Transcription factors are throught to recognize their targets primarily through their sequence-specific binding affinity. A typical factor in yeast binds to short, six to ten base pair sites in promoters [23], with the strength of this binding depending on the specific sequence of the site [25,80]. Both strongly- and weakly-bound sites can impact a gene's expression [81,82], giving most factors a diverse repertoire of potential binding sites to recognize. Accurately describing these repertoires is clearly a key step towards our understanding of target differentiation.

Early efforts towards this goal focused on consensus sequences: discovering a transcription factor's most strongly-bound sequence as well as the positions in this sequence that often vary in sites that are also bound and functional. This representation is computationally tractable but is unable to describe weak sites or well differentiate the relative strength of strong sites. Consensus sequences are still in wide use today, but the currently most popular tool for describing binding sites is the position weight matrix (PWM). These matrices describe the frequency that each base occurs at each position in bound sites compared to that base's frequency in a 'background' model of unbound sequences, converting any given sequence into an often unique score. PWMs can recognize weak sites and theoretical work has also shown the generated score to be a maximum likelihood estimate of the biochemical binding affinity of the sequence to the factor [25].

While PWMs improve upon consensus sequences, in general they continue to poorly identify a given factor's targets. For many factors, due to the short length of their binding sites, even the strongest-binding sequences appear at appreciable frequencies in non-target promoters, and potentially relevant weak sites can be found in most promoters in the genome. Clearly, transcription factors rely on more information than is described in PWMs to differentiate their targets. This information could plausibly be found within the site, and repeated efforts have focused on discovering previously unmodeled dependencies between positions [32,33,83,35]. These efforts have met with mixed success, and currently do not appear to be able to account for the missing information. Alternately, this information could be found in the context of true target sites.

This context can take several forms. Several factors have been shown to have functional mechanisms that naturally specify how their location, orientation, and/or density impacts their binding and effect on expression. Rap1's activation activity was shown to be markedly different depending on which strand its sites were placed and whether or not they appeared as a tandem pair [84, 85]. Reb1 and Abf1 play critical roles in the creation and positioning of nucleosome free regions [86], which are precisely positioned with respect to the transcription start site [87]. This role suggests that, in turn, Reb1 and Abf1 binding sites must be precisely placed in order to function. Other proteins may be less precisely spaced: the homologous factors Met31 and Met32 bind DNA but have no intrinsic ability to activate transcription; their role is to recruit the co-activator Met4 to this sequence [88]. This indirect interaction with the basal transcription machinery may allow these sites to be more flexibly positioned. Finally, beginning with

experiments using artificial constructs [89], cooperativity driven by binding site density has long been thought to play a role in promoter recognition: if the relationship between site number and expression effect is nonlinear, then spurious single sites can be made inconsequential. Many transcription factors, such as Rap1 discussed above, have been shown to bind as dimers. Other factors, Rtg1 and, in A. nidulans, AlcR, bind as monomers but, notably, only recognize and affect expression in promoters with a sufficiently high number of binding sites [90, 91, 92]. Cooperative effects in these cases could be driven by less precise protein-protein interactions or indirectly, through competition with nucleosomes [93, 94]. Taken together, these phenomena could create a promoter-recognition 'signature' for the factor that would render many non-target binding sites irrelevant for regulation and increase the information available for correct target promoter recognition.

Relatively few transcription factors' binding is understood to this level of machanistic detail, but several recent computational works have suggested that these promoter recognition signatures could be a common property. Elemento and Tavazoie [95] used a mutual information approach to simultaneously discover expression-influencing consensus sequences and their location and strand biases, showing that, for a large fraction of the consensus sequences they uncovered, location and often strand informed expression. Following up this work in a larger set of factors, Westholm et al [96] found that the location and strand of many consensus sequences are distributed non-randomly. However, all of these studies have focused on a relatively small number of factors, used consensus sequences instead of position weight matrices, potentially ignoring the effect of weak sites, and, importantly, focused on the properties of individual sites rather than whole promoters, disregarding the effect of site density.

Here we develop an integrated statistical model of promoter signatures for a wide variety of transcription factors. We are able to incorporate and discover factor-specific biases in site location, strand bias, and density. Using this model, we are able to show that spatial information can, in principle, fully compensate for weakly-defined individual binding sites. We validate this information's target differentiation ability using expression changes in transcription factor deletion strains, showing that its target predictions are for most factors better correlated with expression change than are predictions from binding site strength, a thermodynamic model, and, for an appreciable fraction, ChIP-chip.

Results

Description of the model

We use a hidden Markov model to describe the positions of binding sites for a single factor within a set of promoters (fig. 1). For each promoter, a single binary 'regulation' (R) state determines whether or not the emitted sequence will carry the factor's promoter signature. A set of hidden 'site' (S) states generate the observed nucleotide (N) states, one per position in the promoter, according to either a background nucleotide distribution or the appropriate position-specific distribution found within the factor's binding site. A 'consistency' (C) state generated by the last S state ensures that the model generates at least one binding site in each promoter carrying the factor's promoter signature. The model incorporates five parameters: rho, estimating the fraction of sequences in the training set that carry the factor's signature, mu and omega, estimating the center and width of an enriched region of the factor's binding sites, tau, estimating these sites' strand bias, and lambda, a rate parameter which describes the density of sites. As this

poisson-like parameter cannot easily describe the plausible case in which a transcription factor relies on a single binding site for recognition, we also train a similar model which generates strictly one site per promoter. We formally describe these models, as well as fitting and selection, in the methods section.

These models have several valuable properties. They can take advantage of position weight matrices rather than consensus sequences, and, while remaining computationally tractable, they are able to integrate over strong and weak binding sites. Perhaps their key property is their agnostic treatment of the shape of the binding site distribution. As the true shape of the spatial distribution of binding sites is unknown, and may differ between factors [97], we chose to use a conservative flat distribution, creating a plateau-like region enriched for binding sites.

Transcription factors show heterogeneous promoter recognition signatures

We used the Harbison et al [44] ChIP-chip and position weight matrices from the MacIsaac et al [66] analysis as the basis for much of our model training. We screened the binding data in four ways. First, as the position of transcription factor binding sites is much more strongly related to the transcription start site than to the translation start site [98], we removed 5' untranslated regions from our data. Second, we only used intergenic regions containing highly-conserved binding events [66] to lower the prevalence of bound but possibly biologically unimportant sites. Third, although we placed a conservative upper limit on the length of the promoter at 1,000 base pairs, ORFs and other annotated functional sequences were replaced by randomly generated background sequence. Finally, we removed divergently transcribed genes: if a binding site has a spatial bias, but we are unable to assign the site unambiguously to one start site or another, we add preventable noise to our data.

We fitted our model to all ChIP-chip sets having at least twenty promoters meeting our criteria (fig. 2A). We confirm [96,95,97] the presence of factor-characteristic spatial biases of binding sites, extending this analysis to a large number of factors. For each factor, we used likelihood ratio tests over a series of nested models to determine the significance of parameters describing the factor's strand and spacing preferences. As describing binding site density requires a slightly different model structure, we determined the significance of this parameter using an information criterion. Although most factors displayed a significant spatial preference, and there appears to be a diversity of such preferences, we wondered whether this diversity of preferences could be artifacts of differences in promoter length. For instance, the typical promoter region bound by Rpn4 is substantially shorter than the average promoter region; even if Rpn4 sites were randomly scattered through this region, we would expect our model to find Rpn4 sites to be significantly spatially restricted. To control for this effect, we trained our model on data sets with scrambled binding site positions but conserved promoter lengths and binding site composition. In most cases, these parameters fit significantly worse, suggesting that spatial restrictions are driven by more than intergenic sequence length.

We used unbound sequences to control for the effects of weakly or incorrectly specified matrices. If a frequency matrix is likely to appear anywhere, perhaps due to a flaw in our representation of background sequence, then our model could associate with that matrix a well-populated but ultimately meaningless spatial signature. We compensated for this property by fitting our model, for each factor, to regions not bound by that factor in any tested condition. If

we were able to discover any putative signature populated to an appreciable level in these data, we consider the original signature suspect and discard it (see Methods). Although this test is conservative, as many factors bind in a condition specific manner, leaving true binding sites unbound in these ChIP data, only a handful of factors' spatial signatures failed this test.

Our method is sensitive, discovering a substantially higher fraction of factor-characteristic spatial patterns of binding sites than has been shown before. To some extent, this is expected, as our method relies on more sensitive frequency matrices rather than consensus sequences and is not handicapped by attempting to discover spatial relationships and sequence affinity simultaneously.

The tested set of factors exhibits a diversity of spatial patterns. Several factors have sites tightly positioned in relation to the transcription start site. Notably, we recover the hypothesized tight spatial constraint of Reb1 and Abf1 (fig. 3A,B). Several other factors, including Cbf1, Rpn4, and members of the Hap2/3/4/5 complex, also appeared to bind according to tight spatial constraints, and we hypothesize that they may operate under similar mechanistic pressure. Other factors, such as Gcn4, bind more broadly (fig. 3C). Most factors' binding sites were found almost up to the start of transcription, but Fhl1 (fig. 3D) was a notable exception. While relatively few factors exhibited a significant strand bias, we recovered the characteristic bias of Rap1 sites.

Our model is the first description of location bias to explicitly account for binding site density. While some factors appear to recognize single sites in promoters, the typical factor appears to bind to multiple. If multiple sites are a functional necessity for a transcription factor's recognition, then we have, immediately, an intuitive means for increasing a transcription factor's promoter specificity.

Spatial information can offset weak binding information

Many eukaryotic transcription factors have binding sites that are short enough, and nonspecific enough, that identical copies of functional sites appear in most non-target promoters. Examining this formally, Wunderlich & Mirny [99] demonstrated that, unlike those in prokaryotes, virtually all transcription factor binding sites in yeast and other eukaryotes do not contain enough information to differentiate their targets from background sequence. They proposed that binding site density could compensate. The information content of a transcription factor's binding sites can be quantified as the Kullback-Liebler (KL) divergence between the distribution of bases found in these sites and a background distribution [100]. This information content has also been used as a metric to compare the specificity of different transcription factors.

We desired to use this rich framework to compare the specificity of our predicted sequence signatures of target recognition and quantify the increase in specificity they provide over binding sites alone. To this end, we developed a means to calculate the KL divergence between each predicted sequence signature and a background distribution not containing any binding sites. For comparison, we created and repeated this calculation in artificial density- and spacing-agnostic signatures containing a single binding site in each promoter. While calculating this metric directly is all but impossible, as it requires summing over all possible promoter sequences, a sampling approximation produced consistent results (data not shown).

As we expect, a signatures specificity is driven in large part by the specificity of the

original binding site (r**2 = .67). As these sites form the building blocks of any spatial model, factors that have well specified binding sites tend, on average, to also have well specified promoter signatures. Even so, there exists considerable variation in matrix specificity given a certain promoter recognition specificity. In figure 4, we illustrate five factors that, while their signatures share approximately the same overall promoter recognition capacity, have substantially different recognition capacities when spacing, strand, and site density are disregarded. Restriction of these properties is thus able to compensate for a weakly specified frequency matrix.

We also note that, for these factors and nearly all others, overall specificity is greatly increased by the addition of promoter recognition signatures. While factors that rely on one site, without a strand bias, such as Rpn4, gain only a modest specificity increase due to their spatial restriction, most appear to rely on several and show an accordingly large increase in specificity. For instance, Msn4's binding site alone carries roughly one nat of information, which, in theory, is only sufficient to differentiate one third of the genome as its targets-- a far larger role than Msn4, or any other transcription factor in yeast, is expected to play. However, its promoter recognition signature carries more than three nats of information, thought to be sufficient to differentiate roughly 250 targets, only slightly larger than the approximately 200 true targets Msn4 is expected to have [101,102,103]. For reasons we elaborate upon in the Discussion, we do not expect most factors to share this match between calculated specificity and true target size. Nevertheless, these promoter signature driven increases in specificity illustrate a route by which transcription factors can identify their targets, and, as we show in the next section, provide a means by which bioinformaticians might do the same.

Promoter recognition signatures predict expression change in factor deletion mutants

To validate our predicted signatures, we investigated how well they could differentiate factor targets genome-wide. To this end, we measured whether promoters that fit these signatures are likely to exhibit expression changes when their target factor is deleted. Hu et al used microarrays to measure genome wide expression changes in transcription factor deletion mutants [104]. Importantly, these expression changes are a mixture of direct and indirect effects: while direct targets of the deleted factor should show some change in expression, so should targets of other transcription factors and regulatory proteins that are impacted, directly or indirectly, by the factor's deletion. As a large expression change of a particular gene can be explained without that gene being a target, simple correlation between target predictions and expression changes is uninformative. We avoided these difficulties by focusing on a rank list of model predictions: if a model largely predicts targets correctly, its top predictions should be correlated with expression change, this relationship deteriorating as more and more promoters are analyzed.

Using this framework we compared the performance of our model against three other means of predicting factor targets. The first and simplest ranked promoters by the score of the highest-scoring single binding site they contained. The second was a thermodynamic model which was able take advantage of the information found in all of the possible sites to rank target promoters. Importantly, this model does not take into account site location and, unlike our model, handles site density only in an additive manner. The third model simply ranked promoters by their ChIP-chip [44] p-value.

Surprisingly, for several factors, our model equalled or outperformed ChIP-chip target predictions. We illustrate four high-performing promoter signatures in figure 5. In other cases, the results were not so easily interpreted, with different methods' scores being most correlated with expression change at different points in the rank list, or, as with Ume6, which is known to serve as either an activator or a repressor in different contexts, the sign of the correlation fluctuating between negative and positive values. To compare the overall performance of these methods for each factor, we calculated which method's metric produced the largest in magnitude statistically significant magnitude of correlation with the expression data at any point in the rank list. Using this measure, where data was available, our model produced the largest correlation for fifteen factors, ChIP-chip thirteen, the thermodynamic model nine, and the single matrix score model three. Spatial recognition signatures are thus a potentially useful bioinformatic tool for the discovery of transcription factor targets.

Discussion

Spatial specificity compensating for poor site quality

Our principal finding is that transcription factors appear to compensate for poor site specificity through the use of well-specified promoter recognition signatures, often including restricted spacing, orientation bias, and, most importantly, multiple binding sites. As has already been discussed [99], the binding affinities of transcription factors in yeast, and in all eukaryotes, do not specify enough information to differentiate their targets from background DNA. It has been hypothesized 89,99] that this handicap could be overcome through the use of multiple binding sites as a recognition signature.

By focusing on the characteristics of whole promoters, and not, as others have, on the characteristics of individual binding sites, we are able to recover this property of binding site density and show it to be a strong determinant of specificity. We also confirm that transcription factors can have characteristic spatial signatures and significantly expand the repertoire of factors known to exhibit them.

We were able to show that some proteins, such as Msn4, appear to specify exactly as much information in their spatial recognition signature as would be required to differentiate their true target size. There are a number of reasons why we do not expect this to be a general property. First, there is no fast and accurate method for determining what a factor's true target size is. The number of regions determined to be bound using ChIP-chip varies over more than order of magnitude depending on the statistical and conservation criteria employed, and disrupting the target factor and searching for affected genes will always recover a mixture of cis and trans effects. Second, our model does not include properties of transcription factors already known in anecdotal cases to increase their specificity, such as association with different bound factors or tight spacing requirements between co-binding dimers. Finally, proteins may dictate more specificity than they need to simply differentiate their targets. There is a relationship between information and affinity in individual binding sites; if this relationship holds across promoters, with highly-specified promoters being bound a greater fraction of the time than weakly-specified ones, then promoter specificity could have intuitive implications in chromatin remodeling, tight repression, and other biological roles requiring high occupancy.

The use of binding vs. coexpression data

We used binding data as our source of training sets because it is most convenient, allowing us to train models for a large number of factors. It has a number of shortcomings. By focusing on the most strongly bound sequences in the genome, as we must when using these data, we may introduce a bias towards recovering strong sites or large numbers of sites. Perhaps more important, by focusing only on where factors bind, we ignore the arguably significant role that spatial signatures may play in the determination of different expression pattens.

The context-specific properties of Rap1 are an illustrative example [84]. Rap1 binding sites are essential for the activation of many genes, including ribosomal protein genes and genes in the glycolytic pathway. It also is involved in gene silencing near telomeres and at the silent mating loci. Upstream of ribosomal genes, a particular pattern of binding, with sites arrayed in tandem on the coding strand, appears to be critical for maximum expression. Upstream of glycolitic pathway genes, Rap1 usually has one binding site, without an orientation bias, located near one or more Gcr1 binding sites and is apparently essential for the binding of Gcr1. In telomeres, Rap1 appears to bind to a slightly different frequency matrix, perhaps brought about by changes in protein conformation. Several other proteins, such as Cbf1, share Rap1's diversity of function and could potentially share its diversity of spatial signatures.

By focusing on binding instead of expression, we sum over all of these spatial signatures and likely reduce our ability to detect any of them. While we recover Rap1's orientation bias upstream of ribosomal proteins, we mistakenly predict this feature to be general. [96] found a greater prevalence of orientation biases of transcription factor binding sites when they used coexpression rather than binding data, suggesting that promoter signatures may be more coherent in coexpression data sets. Although they are more limited, due to our inability to assign many factors to sets of coexpressed genes, the application and analysis of our model's behavior on these sets is a natural next step.

Use of promoter recognition signatures as a tool

Due to its substantial advantages in time and expense, especially in non-model organisms, computational means of predicting transcription factor targets have been a long sought-after goal. We show that our model has improved performance over a simple thermodynamic model and, indeed, in an appreciable fraction of cases, can produce scores that are more predictive of expression change than p-values from the landmark Harbison et al [44] ChIP-chip experiments. Although we cannot address whether low-scoring promoters by our metric are indeed not bound by this method, due to the abundance of trans effects in deletion mutants, we look forward to comparing our models' performance in curated target sets against ChIP-chip and other models.

Methods

Preparation of promoter regions

We downloaded intergenic regions pre-screened for annotated features ('NotFeature.fasta') from the Saccharomyces Genome Database and used the results of [105] to remove the 5' UTRs. Where data was unavailable, we removed the median 5' UTR length from the beginning of the sequence. We trimmed these sequences to a maximum length of 1,003 base pairs, and we added masking 38% GC content sequence to the 5' ends of sequences shorter than 1,003 base pairs. Finally, we discarded upstream regions that were noted in MacIsaac et al [66] to be part of divergent promoters.

Description of model and algorithm implementation

The model is constructed as a directed graph closely related to a standard first-order Hidden Markov Model. There are four classes of variables. Hidden 'S' variables emit observed 'N' nucleotides. There are three 'background' S states in addition to states representing every forward and reverse position in the frequency matrix: B0 never transitions to any other state, while B1N and B1S can transition to frequency matrix states. Binary 'R' variables at the beginning of the S chain determine whether the nucleotide sequence will contain a promoter recognition signature by determining whether the first S state will be B0, B1N, or a frequency matrix state. At the end of a series of frequency matrix states, a B1S ('site Seen') state is emitted unless another matrix is started immediately. An observed consistency 'C' variable at the end of the 'S' variables takes its observed positive value when the trailing S state is either B0 or B1S, having the effect that all sequences emitted with a promoter recognition signature contain at least one binding site. A related model, here termed the 'monosite' model, can only emit frequency matrix states from the B1N state. We term the original model the 'multisite' model.

The nucleotides are emitted according to the frequencies in the given frequency matrix or from a background model weighted by GC content. In all above analysis, GC context was set at .38.

The value of the R state is given by:

$$P(R)=\rho^{\delta(R=1)}(1-\rho)^{\delta(R=0)}$$

Frequency matrices can be emitted in either the forward or reverse orientation according to a parameter tau. The probability of emitting a frequency matrix from either B1S, B1N, or finished series of frequency matrix states is:

$$P(site)=\frac{\lambda}{1+e^{\eta(|p-\mu|-\omega)}}$$

This value is multiplied by tau or 1-tau depending on the orientation of the matrix. This creates a plateau-shaped distribution of binding sites, with mu and omega specifying the center and spread, respectively. eta either smooths or sharpens the boundaries of the plateau and was set to .1 for all described experiments.

We use the EM algorithm to fit the parameters, starting iterations from fifteen different sets of spatial parameters. The expectation is performed using message passing, and maximum likelihood estimates for rho and tau are calculated analytically. We use simulated annealing to optimize lambda, mu, and omega simultaneously. We implemented the algorithm in C using the GNU Scientific Library and, for information and likelihood calculations, the GNU Multiple Precision Arithmetic Library. The implementation is parallelized with MPI but can be run as a single process.

Spacing controls

We spatially scrambled the original training sets in an iterative fashion. First, we duplicated each set to a minimum size of 600 sequences. Then, at each iteration, we picked a random sequence of length between 5 and 75. We checked if this sequence had any binding sites of score zero at its borders, and if it did, we repeated this process until we found a sequence that satisfied this requirement. We then chose another sequence and random of the same length and

repeated this process. Once a matching sequence was found, we traded the two sequences. We repeated this process 100 million times. We fitted the same number of parameters to these models as were fitted to the originals, and we then repeated the original optimization process, constraining the values of mu and omega to the shuffling-derived values. We determined significance using a likelihood ratio test with two degrees of freedom.

Unbound regions were defined as those which had a ChIP-chip binding p-value greater than .5 in every tested condition. For each factor, we assembled 20 sets at random from intergenic regions meeting this criteria, fitting each set starting from 20 different starting points. A factor's signature was discarded if either: (a) finding the maximum trained rho in each set, if the median of these maximums exceeded .15, or (b) any trained rho value across these 400 fittings exceeded the rho value found in the factor's signature.

Information calculation

We used sampling to approximate the KL divergence between our promoter signatures and a simple background model specified only by GC content. The exact formulation of this divergence is specified as:

$$KL = \sum_{\{N\}} P(\{N\}|signature) \log\left(\frac{P(\{N\}|signature)}{P(\{N\}|background)}\right)$$

By sampling from the model, we are able to replace P(mod) by < 1 / N > , below. {S} refers to a sampled promoter.

$$KL \approx \frac{1}{N} \sum_{\{S\}} \log\left(\frac{P(\{S\}|signature)}{P(\{S\}|background)}\right)$$

Recovery of expression change

We compared four methods in their ability to recover the expression changes found in transcription factor deletion mutants (as described in Hu et al [104]). For each method, we ranked all promoters according to the metrics described below, and then performed a Spearman rank correlation test on each set in a descending rank list.

For the matrix method, we ranked intergenic regions by their highest-scoring motif, for the ChIP-chip method, we ranked intergenic regions by the smallest p-value observed across conditions, and for our promoter signature method, we ranked intergenic regions by the expected value of the R state given by the model. While the rho parameter does not affect rank, we calculated the expectations using rho=.5.

The thermodynamic method relied on the framework described by Stormo [25]. We assumed for each factor that the cell contains a single protein competed for by all of the different intergenic regions. We ranked these regions by their probability of being bound by that factor. The probability of any given binding site being bound being:

$$P(S_\alpha \, is \, bound) = \frac{e^{H(b,i)\cdot S_\alpha}}{Z}$$

where Z is the sum of all the affinities found in the set. Thus, the ranking metric, the probability that at least one binding site is bound, is given by:

$$P(Promoter \, bound) = 1 - \prod_{S_\alpha} 1 - P(S_\alpha \, is \, bound)$$

Figure 1. Description of the model. The model is a directed graph described in the text. A binary regulation variable (green) determines whether a set of state variables (yellow) will emit the observed nucleotides (red) according to a GC-content based background model or a promoter signature model containing one or more binding sites. An observed binary consistency state (orange) ensures that every sequence generated by a promoter signature contains at least one binding site (see Methods).

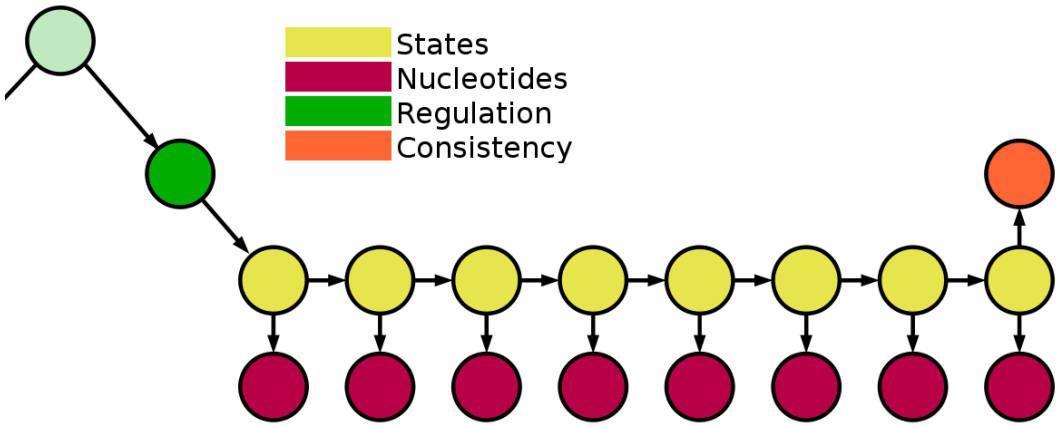


Figure 2. Description of promoter signatures. Promoter signatures for all transcription factors with more than twenty screened bound intergenic regions, excluding those with trainable signatures in unbound regions. Sequence logos depict the frequency matrices described in the main text. The blue enriched region portrays μ plus or minus ω. Regions in gray are those that either failed the shuffling test or did not have statistically significant trained parameter values for μ and ω. The strand column depicts strand bias, from 100% reverse-strand bias (green) to 100% forward-strand bias (red). Circles in the count column depict the expected number of binding sites per promoter. Grey circles correspond to those sequences that better fit the monosite model.

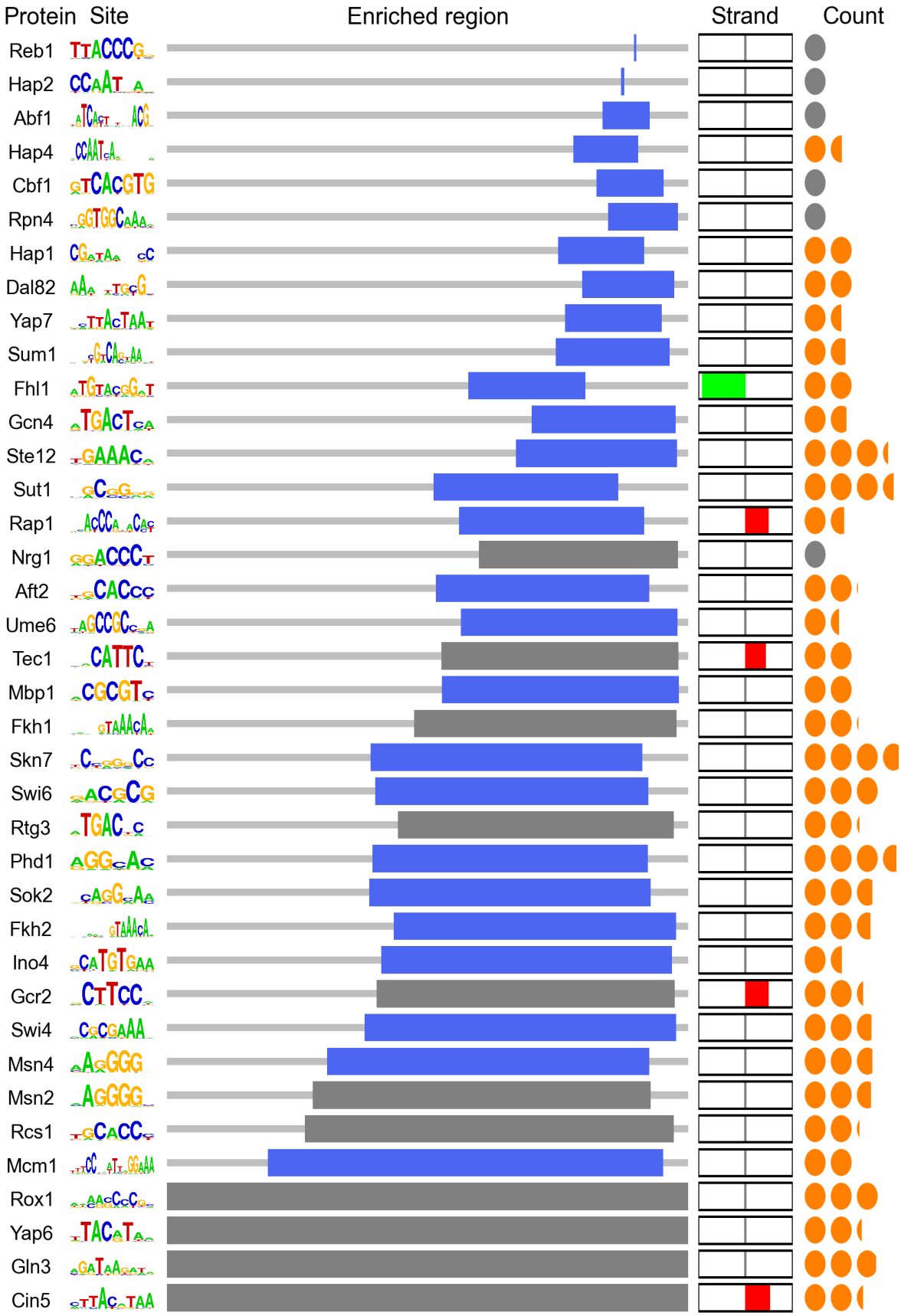

Protein
Site
Enriched region
Strand
Count
Reb1
Hap2
Abf1
Hap4
Cbf1
Rpn4
Hap1
Dal82
Yap7
Sum1
Fhl1
Gcn4
Ste12
Sut1
Rap1
Nrg1
Aft2
Ume6
Tec1
Mbp1
Fkh1
Skn7
Swi6
Rtg3
Phd1
Sok2
Fkh2
Ino4
Gcr2
Swi4
Msn4
Msn2
Rcs1
Mcm1
Rox1
Yap6
Gln3
Cin5

Figure 3. Transcription factors exhibit a diversity of spatial preferences. Score density is plotted against position. Score density is defined as the sum of positive log-two position weight matrix scores in a twenty base window, divided by the total number of possible binding site positions within that window of the training data. Black line is the simulated background score density; grey area is the 95% confidence interval about that line. Confidence intervals are wide in windows far from the transcription start site due to the low number of intergenic regions in the training data reaching this distance. Green area is weighted by the model to be part of the promoter-signature distribution; black area is weighted by the model to be part of the background distribution. Depicted factors are (a) Reb1, (b) Abf1, (c) Gcn4, and (d) Fhl1. No intergenic region used to train Fhl1's spatial signature is as long as 1,000 base pair, creating a blank area.

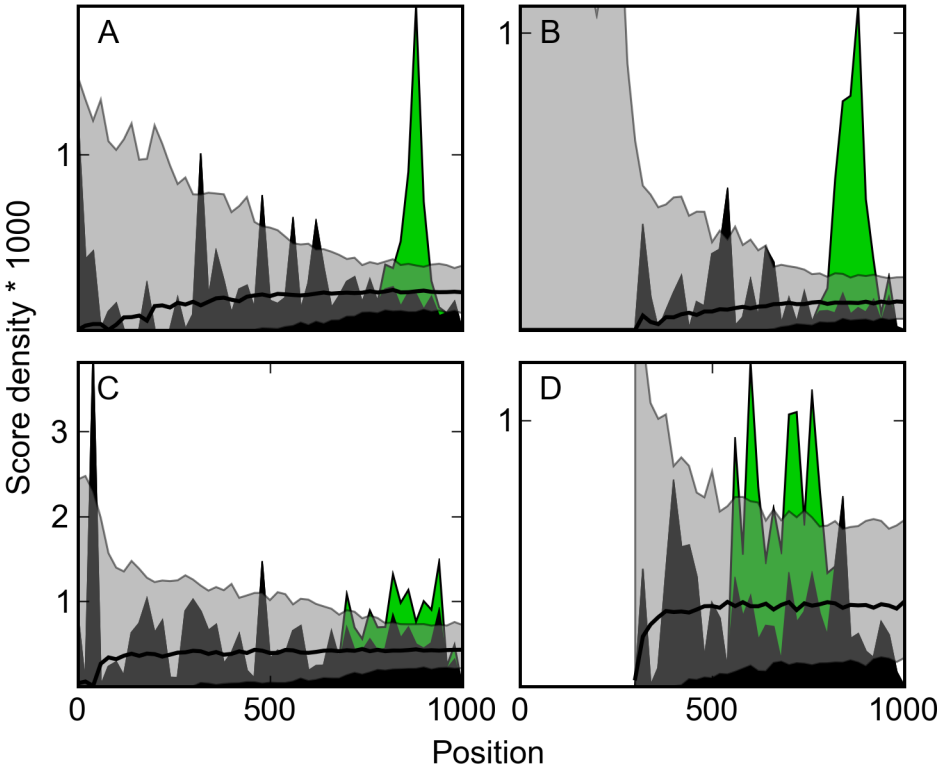


Figure 4. Promoter signatures compensate for and increase the information available to weakly specified binding sites. Information content of promoter signature (orange) and single binding site model (blue) for five transcription factors.

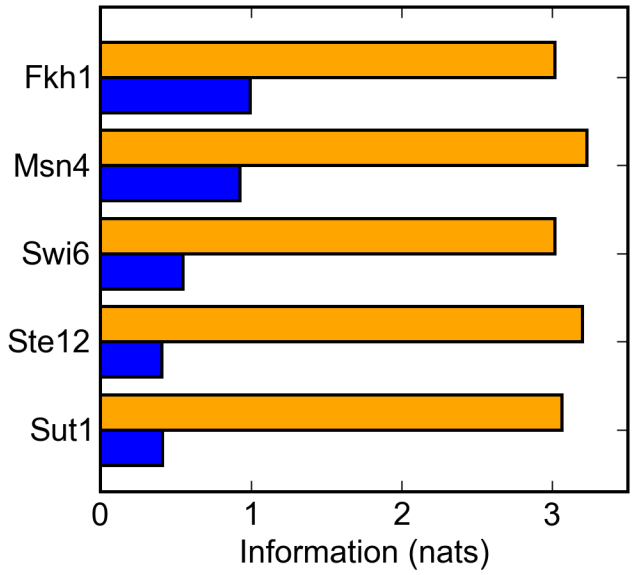
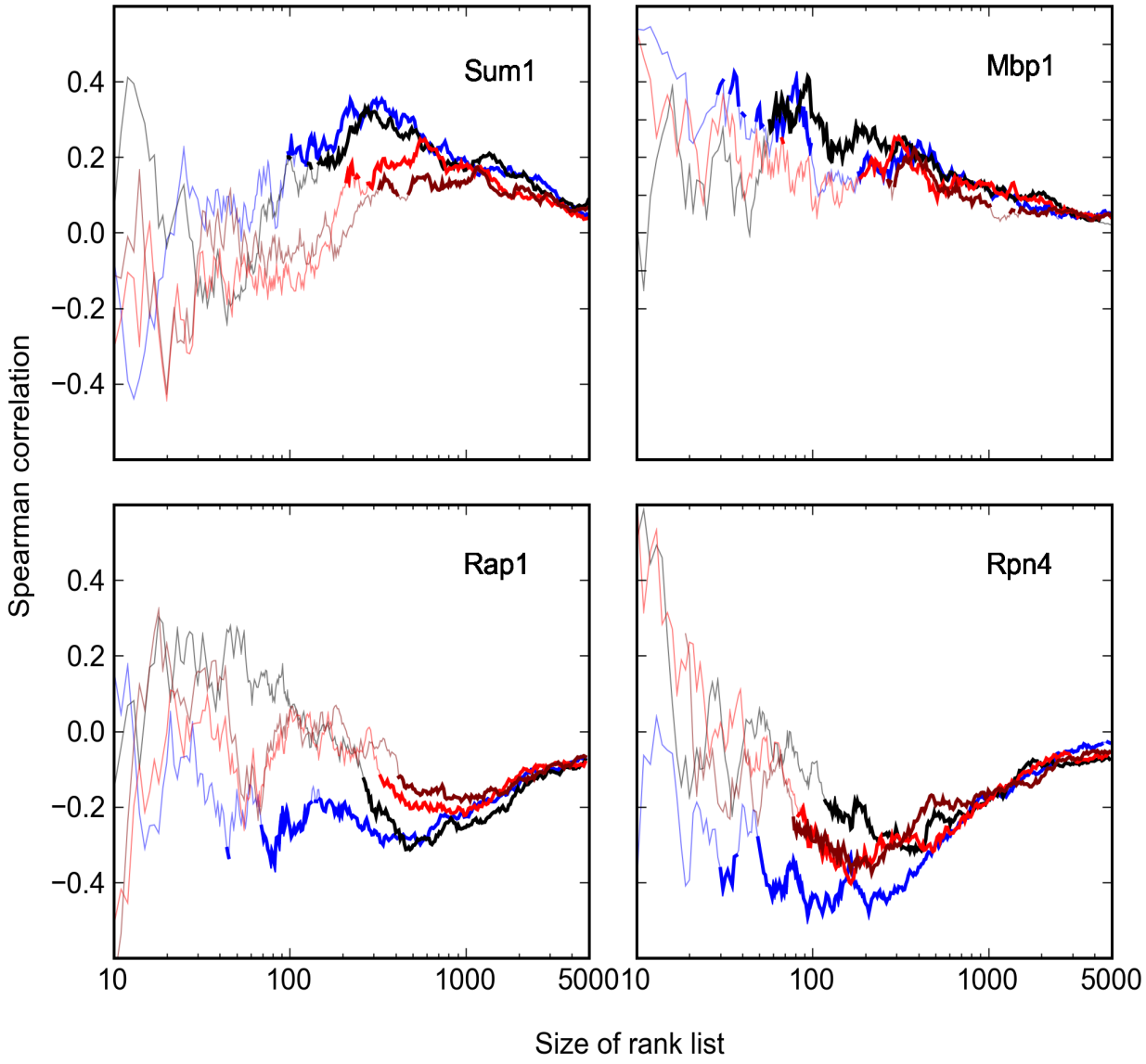


Figure 5. Spatial signature scores are well correlated with expression change in transcription factor deletion mutants. For Rpn4, Sum1, Mbp1, and Rap1, correlation coefficients plotted against data set size by metrics generated by promoter signatures (blue), ChIP-chip (black), maximum matrix score (dark red), and a thermodynamic model (red). For each size N, and metric, the correlation is calculated between the top N promoters by that metric and these promoters' expression change in a deletion mutant. Statistically significant correlations ($p < .05$) are plotted in bold.



Chapter four:

Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers

Abstract

The clustering of transcription factor binding sites in developmental enhancers and the apparent preferential conservation of clustered sites have been widely interpreted as proof that spatially constrained physical interactions between transcription factors are required for regulatory function. However, we show here that selection exclusively on the composition of enhancers, and not their internal structure, can also lead to the accumulation of clustered and overlapping sites with evolutionary dynamics that suggest they are preferentially conserved. We simulated the evolution of idealized enhancers from *Drosophila melanogaster* constrained only to contain a minimum number of binding sites for one or more factors. Under this constraint, mutations that destroy an existing binding site are tolerated only if a compensating site has emerged somewhere else in the enhancer. When the binding specificities of the modeled factors permitted their binding sites to overlap, we observed a significant increase in the evolutionary half-lives and equilibrium density of overlapping sites, primarily because mutations that affect more than one site are accepted far less frequently than those affecting single sites. In our simulations, sites also tended to become closer over time, a result of the strong bias for deletions over insertions in *Drosophila*. The progressive decrease in spacing between sites leads to an overall clustering of sites in the absence of any selection for it, and, because the effect is strongest for the oldest sites, it creates the false impression that proximal sites are more conserved. In simulations of enhancer conservation following speciation, sites tend to be closer together in descendent species than in their common ancestors, violating the common assumption that apparent conservation of a feature in existing species reflects its ancestral state. Finally, we show that selection on binding site composition alone can recapitulate the observed number of overlapping and closely neighboring sites in real *D. melanogaster* enhancers. Thus this study calls into question the common practice of inferring "*cis*-regulatory grammars" from the organization and evolutionary dynamics of binding sites in developmental enhancers.

Introduction

The transcriptional output of developmental enhancers is affected by the spatial organization of the transcription factor binding sites they contain. The relative positioning of sites is known from individual cases to modulate direct competition between factors for the same site [20,106], cooperative and repressive interactions between transcription factors [107,108], and the formation of higher order regulatory complexes [11,109,110]. However, we have a precise understanding of the relationship between binding site organization and function for few, if any, developmental enhancers.

In the absence of efficient experimental protocols for dissecting enhancer function, recent efforts have attempted to infer functional constraints on binding site organization from the distribution and evolution of binding sites in enhancers of interest. For example, we examined developmental enhancers in species distantly related to *D. melanogaster* and found a strong preferential conservation of overlapping and proximal sites [60], a result which was confirmed by a recent survey of enhancer evolution across the twelve sequenced *Drosophila* genomes [38]. Others have focused on the density of overlapping and proximal sites, finding that both are significantly enriched [39,40]. All of these studies, including ours, reached a similar conclusion: that the evolutionary dynamics of binding sites in developmental enhancers suggest that clustered and/or overlapping sites are common functional necessities for enhancer activity.

This shared conclusion was premised on the idea that the observed non-random arrangement of sites must be a result of selection on the relative positioning of sites within enhancers. However, alternative explanations for these phenomena, especially the possibility that such arrangements might arise as a byproduct of other mutational and selective pressures [111], have not been explored. We were interested, in particular, in how selection to maintain the composition of enhancers might affect the distribution of binding sites within them.

Here we simulate the evolution of real and synthetic *D. melanogaster* enhancers constrained only to maintain their binding site composition, and investigate the spatial organization of binding sites within enhancers evolving with no direct selection on the arrangement of sites within the enhancer. We show that a simple global constraint on enhancer composition is sufficient to produce many of the organizational and evolutionary features observed in real enhancers, including enrichment and apparent conservation of overlapping and clustered sites.

Results

Simulating enhancer evolution

To explore the properties of enhancers evolving under selection on binding site composition, we created synthetic enhancers in which a predefined number of binding sites for one or more transcription factors were randomly positioned in a background of randomly generated sequence with the same composition as *D. melanogaster* non-coding DNA. We subjected these synthetic enhancers to random mutations sampled from the distribution of substitutions, insertions and deletions observed in *D. melanogaster* [112]. If the number of sites in the entire enhancer fell below a specified threshold, we rejected the new sequence. Otherwise, it was carried through to the next mutational step (Figure 1).

We compared the behavior of this model of the evolution of a single enhancer with a strict fitness cutoff to simulations of a large population of enhancers in which suboptimal

sequences were assigned a fitness penalty rather than being immediately removed. None of the measures of binding site distribution and evolution discussed below differed appreciably between these models (see Appendix A).

Since these population simulations required significantly greater computational resources, we present only the results of the simpler model below.

Binding site turnover

The most basic property of our model of enhancer evolution is that most mutations that destroy a binding site will be rejected, as they bring the number of sites present in the enhancer below the specified fitness threshold (Figure 1B). However, the small size of most binding sites means that they are generated *de novo* by random mutation at an appreciable rate. And, once new sites are generated, mutations that destroy existing sites will be tolerated (Figure 1C), leading to non-homologous site conservation, or "binding site turnover" [56,58,59,113].

The rate of turnover of different factors in real enhancers is not the same. To examine the extent to which this rate variation reflects inherent properties of the turnover process itself, and not differential selection on binding site positions, we simulated the evolution of enhancers constrained only to have a single site matching real, or randomly generated, transcription factor specificities. The rate of turnover varied considerably, depending on the size of the recognition site, its base composition and degeneracy (Figure 2), with the variance primarily due to variation in the rate at which new binding sites are generated from random DNA. Since high-information sites are generated from random sequence at a lower rate, they turn over more slowly.

The expected half-life (measured in mutational distance) of Bicoid (BCD) and Krüppel (KR), two typical *D. melanogaster* transcription factors were between one and two substitution per site, or around 50 to 100 million years. This is consistent with published estimates of the turnover rates for functional sites in real enhancers, which has been estimated to be around one to two turnover events per site per hundred million years [59,114]. We found other factors, such as Hunchback (HB) and Giant (GT), to have a shorter half-life than expected in our simulations, due to autocorrelations in their matrices not usually found in the randomized matrices.

Selection on binding site composition alone leads to conserved structure in enhancers

Some transcription factors overlap in their binding specificities, such that the same bases can be parts of binding sites with multiple factors. For example, BCD and KR have overlapping specificities [115-117], and in specific cases competition between them for overlapping sites plays an important role in producing specific expression patterns [118,119]. The high frequency of overlapping BCD and KR sites in other embryonic enhancers has been used as evidence for the generality of this mechanism [39].

However, when we simulated the evolution of synthetic 1,000 bp enhancers constrained to contain five BCD and five KR sites in 1,000 base pairs, we find an almost twofold elevation in the frequency of overlapping BCD and KR sites compared to the random expectation BCDKR (Figure 3A). Thus selection acting to preserve enhancer composition alone indirectly leads to "higher order" structure in enhancers. This phenomenon is not specific to BCD and KR, rather it is a general property of factors with overlapping binding specificities (data not shown).

The increase in the density of overlapping sites is almost entirely due to their increased half-life relative to isolated sites. In the BCD/ KR simulations described above, which had no

explicit selection to maintain overlapping sites, overlapping sites persisted 1.5 to 2.0 times longer (depending on the specific choice of matrix) than isolated sites (Figures 3B, S1-2).

This difference in half-life between overlapping and isolated sites not only increases the density of overlapping sites, it significantly alters how they are classified in comparative genomic analyses. Their longer half-life means that overlapping sites are more likely to be found at orthologous positions in related species. In particular, at evolutionary distances in the range typically used for comparative analyses (around one substitution per site) the likelihood of finding an orthologous overlapping pair of BCD and KR sites is two times larger than the likelihood of finding an orthologous singleton site (Figures 3B, S1 and S2).

Thus, our simulations show that selection to maintain enhancer composition not only leads to an increase in the density of overlapping sites, it also makes it appear that selection is acting to specifically preserve them.

A deletion bias induces conserved binding site clustering

Binding sites in real enhancers are clustered, with an excess of short inter-binding-site distances at the expense of long ones [39,40]. This clustering has been interpreted as evidence that long-range interactions between transcription factors or between transcription factors and nucleosomes are required for proper gene regulation [39,40].

However, in our simulations, we also observed an increase in the proportion of small spacers (Figure 4A). This induced binding site clustering occurred whenever the mutation model included a bias for deletions over insertions, a known property of *Drosophila* species [120]. When simulations were run with only point mutations, or with balanced insertions and deletions, no increase in short spacers was observed.

Unlike point mutations, deletions can disrupt multiple non-overlapping binding sites. In our simulations, deletions affecting two or more sites were less than half as likely to be accepted as deletions affecting single sites (10.5% compared to 23.2% of the time). Thus it is possible that the induced binding site clustering arises from the protective effect proximal sites have against each other's deletion (Figure 4B). Indeed, in simulations that exclusively involved deletions, tightly-spaced but non-overlapping sites showed a substantial increase in half-life (Figure S3). However, in simulations with a realistic balance of mutations and indels this effect was minimal (Figure S4), as the frequency of multi-site deletions was low relative to single site deletions and point mutations.

Instead, the induced binding site clustering appears to be driven simply by the deletion of spacer DNA between sites. Since, in our simulations, deletions between sites occur more frequently than sites are lost, sites get closer together over time, distorting the distribution of inter-site distances. A corollary of this phenomenon is that sites that are observed to be close together tend to be older, and therefore more likely to be labeled as conserved, than isolated sites (Figure 4C). Thus, both binding site clustering and an apparent preferential conservation of clustered sites are expected to occur even in the absence of any selection on enhancer organization.

A deletion bias distorts evolutionary inference

Sequence features is present in multiple related species are generally considered to reflect those found in the shared ancestor, whether through selection or simply common descent.

However, the deletion bias induced tendency for sites to get closer together over time distorts this relationship. To illustrate this, we placed two sites at a fixed distance and monitored the distance between sites over time in a large number of independent simulations. With indels, but no bias towards deletions either in frequency or in average length, the intersite spacing quickly diverges between simulations (Figure 5A). However, with the observed *Drosophila* deletion bias, the spacing between sites in the different simulations is strongly correlated (Figure 5B). Thus, with a deletion bias, the spacing between sites after speciation will appear conserved and yet not reflect either selection or the ancestral state.

To examine how this relationship between inter-site spacing and age might affect evolutionary inference, we simulated the divergence after speciation of regulatory sequences containing pairs of binding sites separated by varying distances. We then compared, at different times after divergence, the inter-site spacing in orthologous evolved sequences. For each of these cases, at varying times after divergence, we performed a simple test of spacing conservation, assessing whether both orthologous pairs of binding sites met an arbitrary spacing criterion. If both did, their spacing was considered to appear conserved. We observed that, with or without a deletion bias, this result often appeared by chance (fig. 5C). However, incorporation of a deletion bias substantially increased this misleading appearance of conservation, and at longer timescales, an appreciable fraction of sites that were distantly separated in the ancestor appeared to share a conserved close spacing in the descendants (fig. 5D,E).

A plausible evolutionary scenario explains positional information in a Drosophila enhancer

To assess whether the above-described effects could replicate the degree of binding site overlap and clustering that is observed in extant enhancers, we simulated the evolution of the well-characterized *eve* stripe 2 enhancer [119]. with compositional constraints derived from the extensive biochemical and genetic literature on this enhnacer. In particular we required five KR, 10 BCD, three HB, and five GT sites [121], and a single Zelda [122] binding site (see Table 1). We also required that a certain number of sites for each factor be predicted high-affinity sites (based on the number of high-affinity sites in the *D. melanogaster* enhancer).

We simulated 1,000 replicates of this enhancer to twenty substitutions per site, and found that both the number of overlapping BCD and KR sites, and the number of sites in close proximity to others, in the real enhancer were well within the range typically generated by this architecture-free evolutionary model (Figure 6A,B).

Discussion

New molecular methods and ever more sophisticated computational approaches have made significant progress towards understanding the mechanisms of gene regulation. Sequence affinities and binding sites for many transcription factors in many organisms are known, and increasing attention is now being paid to the 'grammar' that may link them together [81,123,124].

A common strategy in our work and that of many of our colleagues has been to infer functional constraints on enhancer activity from the apparent conservation of aspects of the organization of transcription factor binding sites within enhancers. However, the results of the simulations presented here show that many of our conclusions were based on naïve assumptions about the expected distribution of binding sites in enhancers evolving with no constraints on their organization.

The value of simulations

In retrospect, the properties we observed are straightforward consequences of coupling selection on binding site composition with a deletion biased mutational process. One does not need simulations to see why overlapping sites will clearly turn over less frequently than isolated sites, that a deletion bias will drive sites closer together over time, and how both phenomena distort comparative analyses.

But as self-evident as these results may appear, they have never been noted before, despite more than a decade of intense comparative genomic analysis of enhancer structure and function in *Drosophila*. Indeed prior to performing these simulations we did not consider that the clustering of binding sites in *Drosophila* enhancers might arise from a deletion bias. We simply attempted to have our simulations accurately reflect the underlying mutational process in our simulations, with the consequences evident only in the results. This highlights the value of the simulations of simple evolutionary processes in uncovering unappreciated consequences of our models and assumptions.

Furthermore, while the general effects of selection on binding site composition and of a deletion bias can be intuited, specific quantitative aspects of the model are difficult to work out analytically. For example, while we have developed a mathematical model for the effect on half-life of overlapping sites in enhancers (see Appendix B), it is difficult to extend this model to enhancers with multiple sites. Simulations can answer these questions simply and effectively.

Generality

The simulations we performed here used non-coding DNA, transcription factor binding sites and mutation patterns from *Drosophila*. Interspecies differences in the composition of non-coding DNA, specificity of transcription factors and base substitution patterns will have minimal effect on our conclusions. But differences in the indel rate and the balance of insertions and deletions could significantly alter the existence or magnitude of the induced binding site clustering. Although the deletion biased mutation process we used in our model is often thought of as a *Drosophila* specific phenomena, there is increasing evidence that short indels are deletion biased in all species [125-130]. Thus, we expect this effect to be general, although the magnitude will differ depending on the indel rate and bias (see Appendix C).

Conclusions

Lynch has eloquently argued that biologists are often too quick to assume that organismal and genomic complexity must arise from selection for complex structures, and too slow to adopt non-adaptive hypotheses [111]. Our results lend additional support to this view, and extend it to show that indirect and non-adaptive forces can not only produce structure, but also create an illusion that this structure is being conserved.

We do not doubt that many aspects of transcriptional regulation constrain the location of transcription factor binding sites within enhancers. Indeed a large body of experimental evidence supports this notion, and we remain committed to identifying and characterizing these constraints. But if this process is to be fueled by comparative sequence analysis, as we believe it must be, it is essential that we give careful consideration to the neutral and indirect forces that we now know can produce evolutionary mirages of structure and function.

Methods

Simulation of enhancer evolution

Starting sequences 1,000 basepairs in length were generated randomly to match the base composition of *D. melanogaster* non-coding DNA, and binding sites were added to bring the starting density of sites to the specified thresholds. Mutations were sampled randomly from point mutations, insertions, and deletions. 80% of mutations were point mutations generated from an HKY85 [36] model with GC content 40% and kappa two; 12% were deletions and 8% insertions with size distributions drawn from [112]. The deletion bias (60%), and proportion of all mutations that were indels (20%), were also according to [112]. Except where noted, simulations took place for 100,000 mutation/selection rounds. Insertions and deletions inside this sequence respectively removed or added base pairs at the nearest edge of the sequence. All new base pairs added were drawn from a 40% GC content random pool of bases. The simulation software was written in Python and utilizes the Motility [131] binding site identification package.

In the simulations in presented in figure five, we sought only to examine the evolution of site spacing over time and not the conservation and/or turnover of individual binding sites. Thus, we preconditioned in each case that neither could binding sites be generated from random sequence nor could existing binding sites be disrupted. To this end, in these simulations, all mutations affecting positions contained within existing binding sites were considered precluded by selection and discarded, and, similarly, the sequence was not scored for new binding sites created by mutations. We generated figures A-B and C-E, respectively, by simulating 980,000 and 480,000 300 base pair sequences to thirty and ten substitutions per site. In the even indels case, the distribution of insertion lengths was set equal to the distribution of deletion lengths.

Simulations using BCD and KR used matrices from *in vitro* footprinting [132], one-hybrid assays [133], and SELEX [46], with cutoff scores chosen to match expected numbers of their sites in the *even-skipped* stripe two enhancer: 5.5, 4.9, and 4.1 for BCD and 5.6, 4.1, and 0.0 for KR for the three sources of matrixes. GT and HB matrices were taken from footprinting and were both required to meet a score cutoff of 4.9. Unless noted otherwise, simulations used matrices from the footprinting data set.

Properties of the simulations were computed following a lengthy (~30 subs/site) burn-in period that allowed the randomly generated starting model to reach equilbrium. We tested several sets of neutral mutation and selective parameters to make sure this burn-in period was sufficient (fig. S5-7).

Generation of randomized binding sites

We chose binding site lengths randomly between five and twelve. At each position, we chose a consensus nucleotide and assigned its frequency by sampling a Gaussian with mean 0.8 and standard deviation 0.2. Subsequent nucleotide frequencies were chosen similarly, each being given a frequency chosen from a Gaussian with a mean and standard deviation of 80% and 20% of the remaining probability mass, respectively. Weight matrices were constructed against a 40% GC bias and threshold scores were sampled from a uniform distribution spanning zero to the maximum scores of the sites. Information content was calculated by weighting all N-mers above the score threshold with the GC bias and subtracting the information in an N-mer of random sequence of equal length and GC bias.

Conditional probability of overlapped sites

To find the expected probability KR and BCD sites would overlap in random DNA, we sampled random ten-mers from a 40% GC background distribution. If this sequence contained a KR site, then we added flanking sequence of length N-1, where N is the length of a BCD site. If this sequence also contained a BCD site, then we considered it as an overlap. The probability of a BCD site generating a KR site was found in an analogous manner. The post-selection conditional probability was directly calculated by simulating an enhancer with five sites for each transcription factor as described above and counting observations of singleton and overlapped binding sites.

Half-lives of binding sites

We determined the half-lives of sets of binding sites by randomly sampling individual sites in our simulations and observing their degradation as the simulations progressed. Our data consisted of simulations of 1,000 enhancers, each run for 30,000 iterations. For each enhancer, after a burn-in period of 10,000 iterations, we took a 'snapshot' of the binding sites present every 3,000 iterations. In each subsequent iteration of the simulation, the presence or absence of each binding site in the snapshot was assessed: if it had been destroyed by a point mutation or indel in that iteration, then a site 'death' was recorded. This process was repeated for 2,000 post-snapshot iterations of the simulation.

Generation of the even-skipped stripe two enhancer

We used one-hybrid binding sequences for Hunchback, Giant, BCD, and KR from [39] and created weighted matrices as described. We used the same methods to generate a Zelda-consensus matrix from the sequences listed in [134]. Our enhancer sequence and matrices can be found at http://rana.lbl.gov/~rlusk/mirage/. In order to determine the required number of sites for each matrix, we assessed the number of hits it had to the *eve* stripe 2 sequence at several score cutoffs. If the number of hits at a given score cutoff exceeded the number expected by chance, then this number/score cutoff pair was accepted as a requirement, provided that it did not substantially increase the total required number of sites for that factor beyond that described in [23]. The constraint on the enhancer is available in the supplementary materials (Table 1).

Acknowledgements

We would like to thank Brant Peterson and members of the Eisen lab for extensive and fruitful discussions. We would also like to thank Leonid Teytelman, Mathilde Paris, Aaron Hechmer, Steven Maere and Matt Davis for comments on the manuscript.

Table 1. Modeled constraint on *eve* stripe 2 enhancer

Factor	**Threshold**	**Count**
CAGGTAG	9.0	1
Giant	5.0	1
Giant	2.5	4
Bicoid	7.0	3
Bicoid	4.5	7
Kruppel	8.0	1
Kruppel	7.0	2
Kruppel	4.0	2
Hunchback	7.0	1
Hunchback	5.0	2

Figure 1. Simulation of enhancers under a compositional constraint

A. We started each simulation with, for example five BCD (red triangles) and five KR (blue circles) binding sites randomly positioned in a randomly generated 1,000 basepair sequence generated with p(A)=p(T)=0.3 and p(C)=p(G)=0.2. Each iteration in the simulation involved a mutation step followed by selection requiring that at least five BCD and five KR sites be present.
B. A deletion (red bar) eliminates a KR site bringing the total number to four, and leading to the rejection of the mutation.
C. A mutation creates a new KR site (bringing the total to six) and is accepted. The subsequent deletion of an original KR site (red bar) does not reduce the total below five and is accepted, leading to a binding site turnover event.
D. Sample run of the simulation over 1,500 mutation-selection rounds. The course of the simulation proceeds from top to bottom, with BCD sites represented in pink and KR sites represented in blue. Overlapped BCD/KR sites are darker and purple.

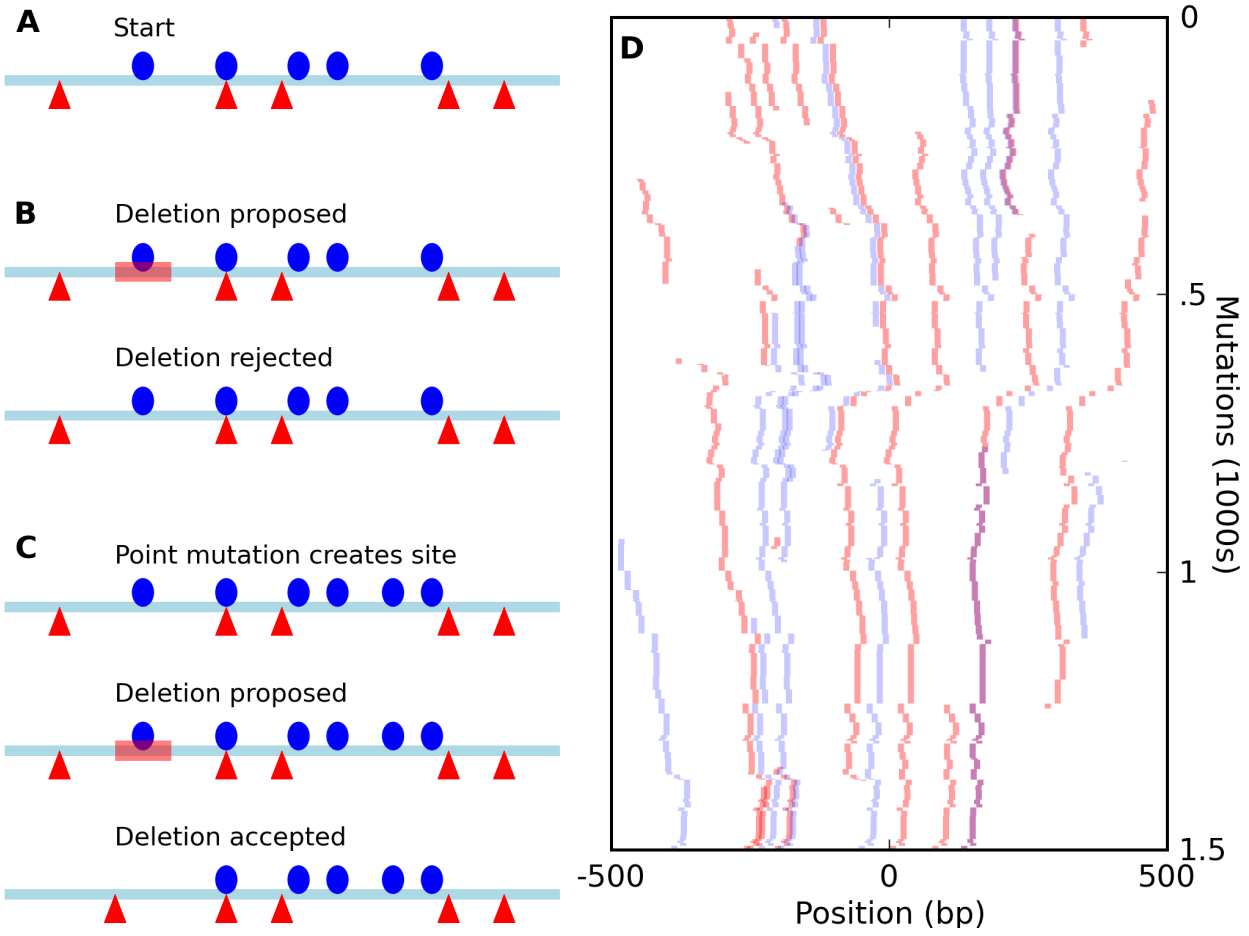


Figure 2. Rate of binding site turnover is correlated with information content

The log of the half-life of different artificial and real binding sites against their specificity. Synthetic binding sites are plotted in gray, while sites derived from *Drosophila* transcription factors are highlighted: KR (blue oval), BCD (red triangle), Giant (GT, green diamond), and Hunchback (HB, cyan hexagon). Specificity is defined as the difference in the information between the binding site and a random sequence of the same length.

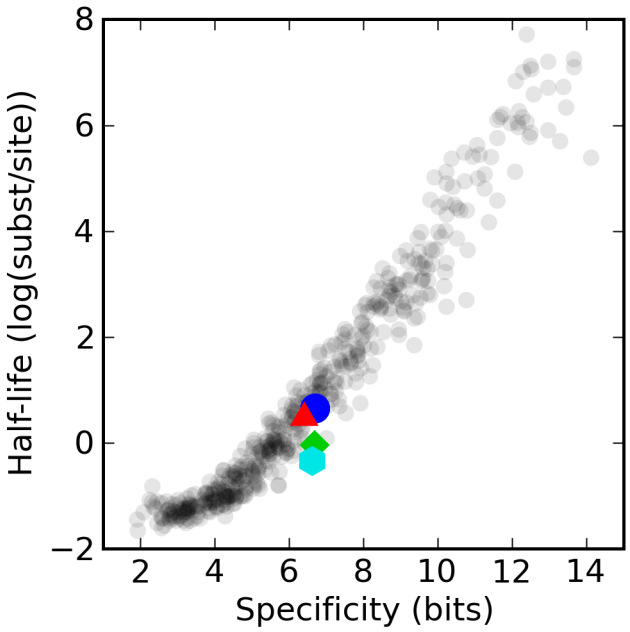


Figure 3. Overlapping binding sites are enriched and "conserved" in simulated sequences

A. The post-simulation (S) probability of observing a KR site conditioned on seeing a BCD site (blue) and a BCD site conditioned on seeing a KR site (red) is significantly higher than the expected probability (E) in random DNA for binding matrices derived from *in vitro* footprinting. **B.** Overlapping sites (solid line) are more likely than isolated sites (dashed line) to persist in simulations at a wide range of mutational distances.

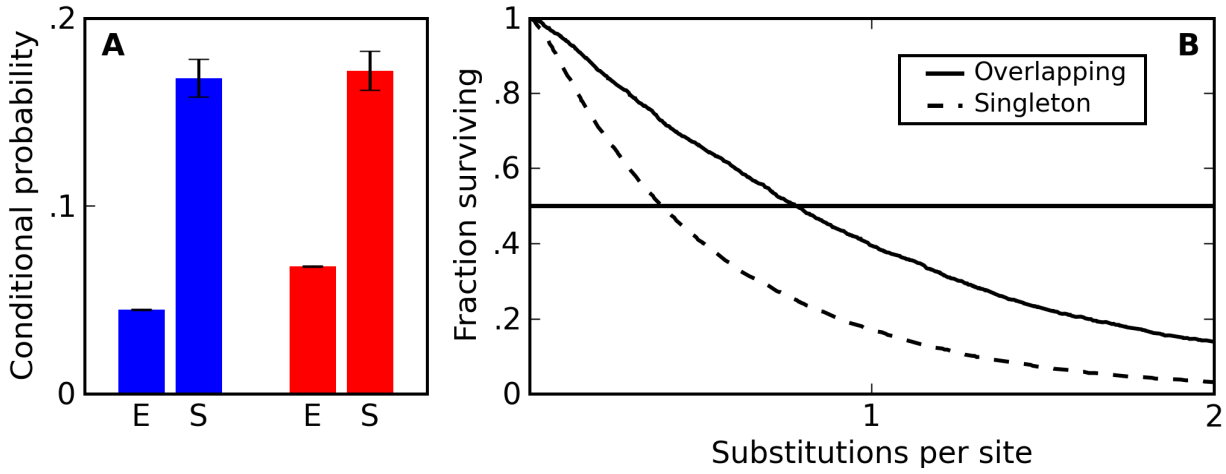


Figure 4. A deletion bias leads to clustering of sites and the apparent conservation of clustered sites

A. The distribution of spacer lengths between binding sites during simulations in which 0% (black), 20% (light green), and 40% (dark green) of mutation events are indels with a 3:2 deletion:insertion bias.
B. The percent probability that a deletion event affecting a given binding site is accepted by our selective process for adjacent sites (Adj; sites that are touching) or far sites (Far; those with a spacer of at least twenty base to the nearest neighboring site).
C. The distribution of the average age of binding sites as a function of their distance to their nearest neighbor shows that clustered sites appear more conserved than isolated sites, even though no such selection was applied in the simulations.

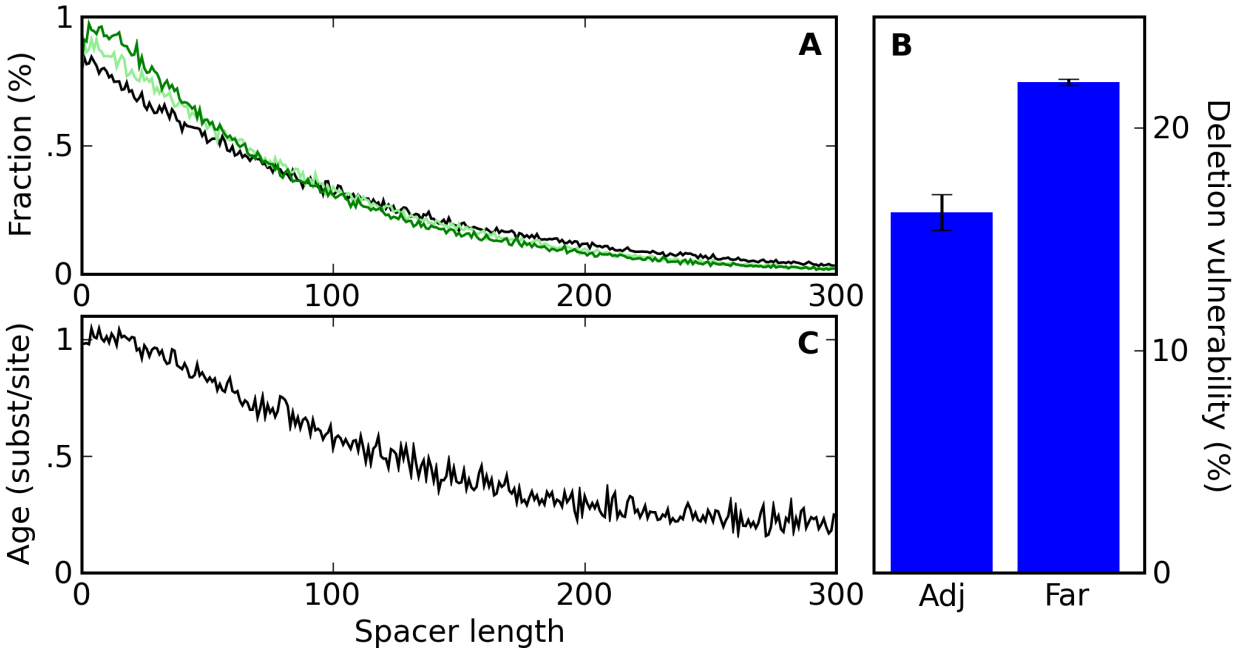


Figure 5. A deletion bias reduces the functional information that can be derived from spatial conservation analysis

A-B. Following an initial starting condition where two binding sites are 100 base pairs apart, the evolution of their spacing is simulated where either (A) there is no bias towards deletions or (B) the distribution of indels approximates that found in *Drosophila*. The probability of observing the sites separated by a given distance after a given number of substitutions is shown on a scale of deep blue (zero) to deep red (≥ 2%). Without a deletion bias, site spacing rapidly becomes unpredictable. However, the deletion bias, on average, ratchets sites together over time, correlating any two pairs' of sites evolution.
C-E. After starting 30 (C), 50 (D), or 100 (E) base pairs apart at a speciation event, orthologous pairs of sites are subjected to a simple test of spacing conservation. If both pairs of sites are separated by a distance of 30 base pairs or less after diverging by a certain number of substitutions, their close spacing is considered 'conserved.' We plot the chance that, given that none of the sites themselves have degraded, this apparent conservation could be created by a neutral model. This neutral model may have a balance of insertions and deletions (blue) or a deletion bias approximating *Drosophila's* (green). When no deletion bias is present, the chance that apparently conserved spacing is explained by neutral forces decreases over time, allowing better discrimination of 'true' conservation via negative selection. *Drosophila*'s neutral mutation pattern not only reverses this trend (C), but also induces a substantial fraction of originally distantly-spaced sites to appear to have a conserved close spacing (D, E).

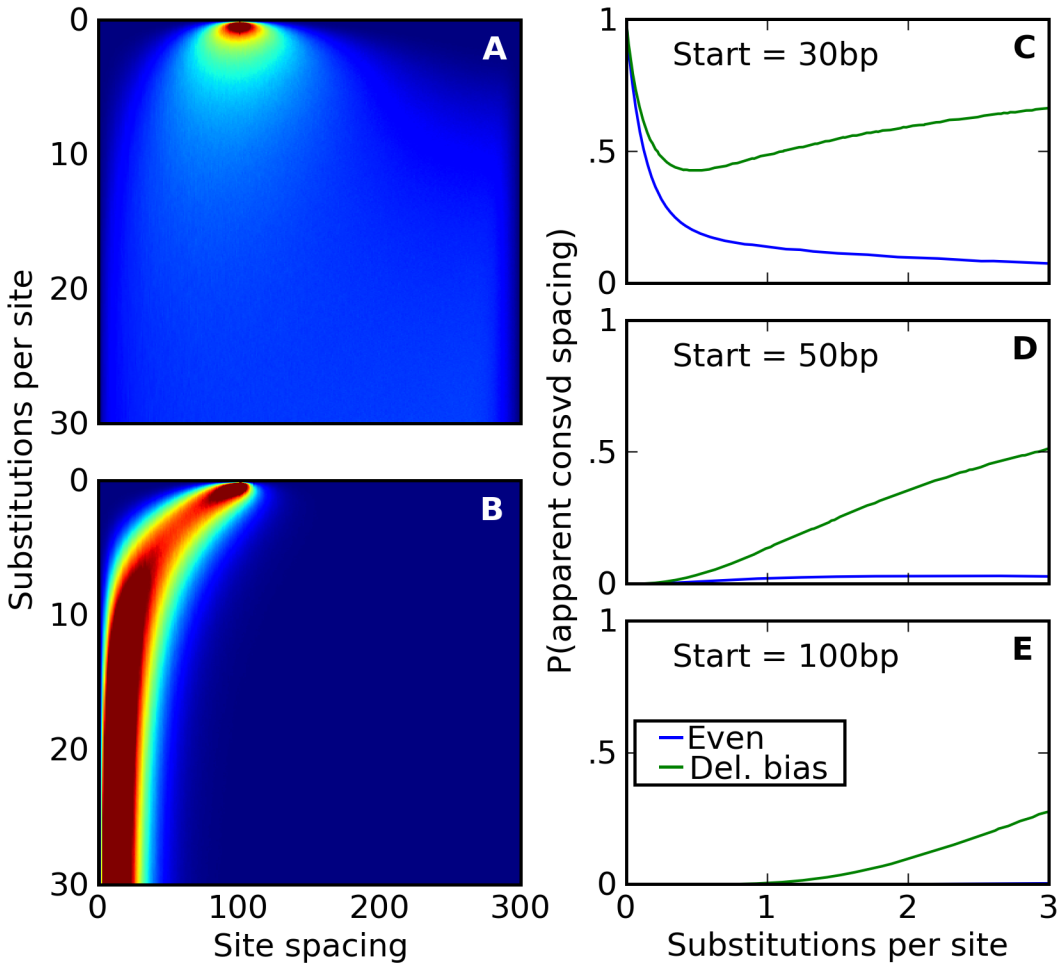

A
B
C
D
E
Substitutions per site
Site spacing
P(apparent consvd spacing)
Substitutions per site
Start = 30bp
Start = 50bp
Start = 100bp
Even
Del. bias
0
10
20
30
100
200
300
1
.5
0
1
2
3

Figure 6. Simulations recover the grammar of the *eve* stripe two enhancer

One thousand simulations of the *eve* stripe 2 enhancer (see Methods) resulted in variable numbers of overlapping BCD and KR sites (A, grey histogram) and sites within 10 basepairs of each other (B, grey histogram). The number of overlapping BCD/KR site pairs, and closely spaced sites in the real *eve* stripe 2 enhancer are shown in red. That the real numbers are comfortably within the range produced by these simulations demonstrates that the higher-order structure in real *D. melanogaster* enhancers could plausibly have arisen solely from deletion biased mutation and selection to maintain binding site composition.

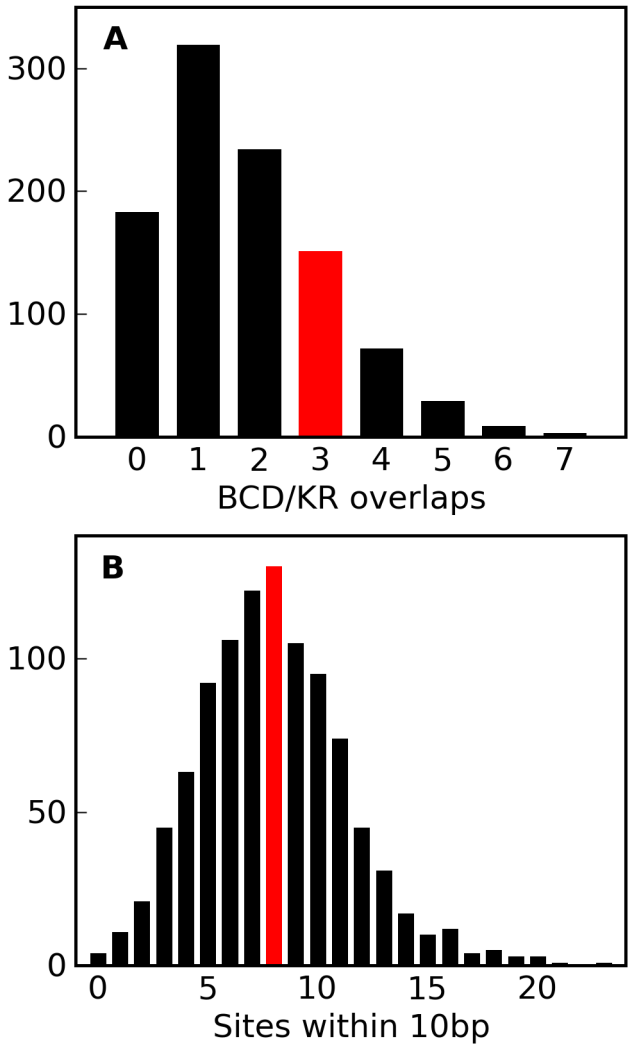


Figure S1. The half-life of overlapping or singleton sites computed using BCD and KR specificity matrixes from one-hybrid data.

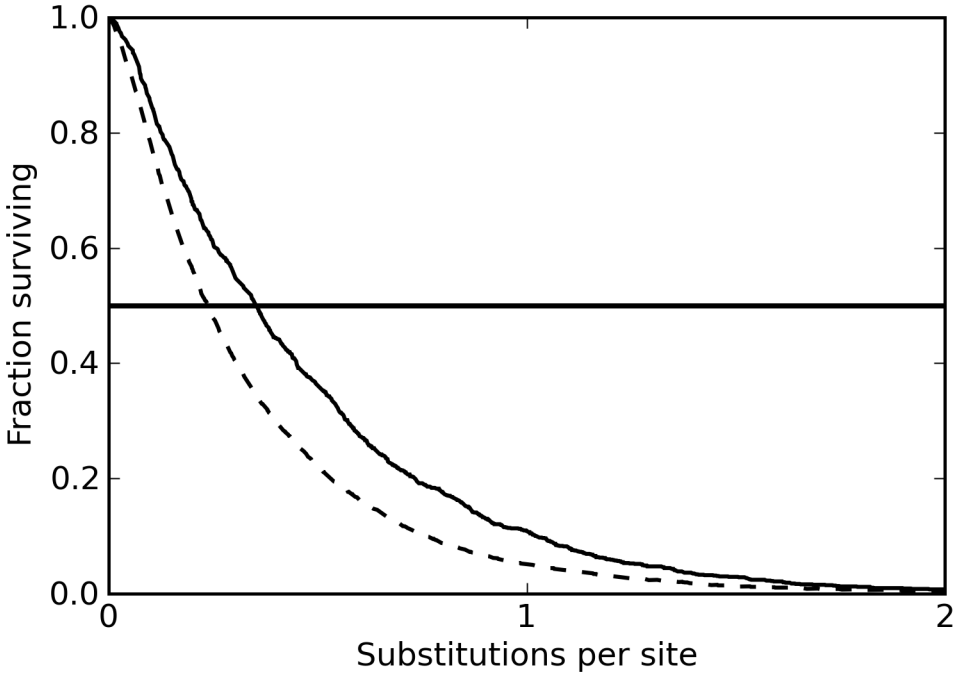


Figure S2. The half-life of overlapping or singleton sites computed using BCD and KR specificity matrixes from SELEX data.

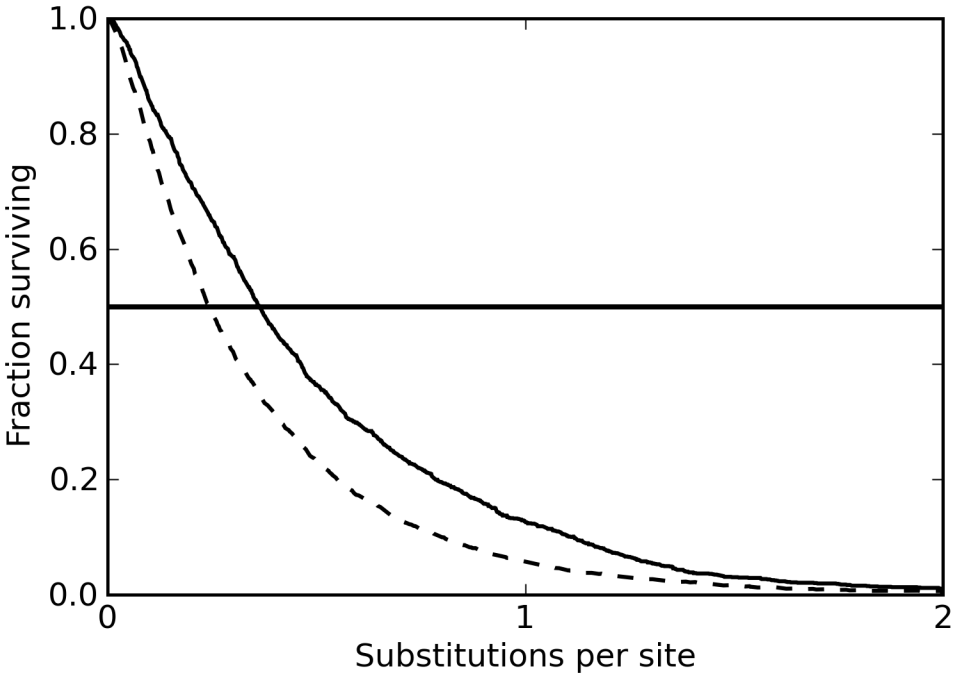


Figure S3. In simulations that exclusively involved deletions, tightly-spaced but non-overlapping sites (solid lines) showed a substantial increase in half-life over isolated sites (dotted lines).

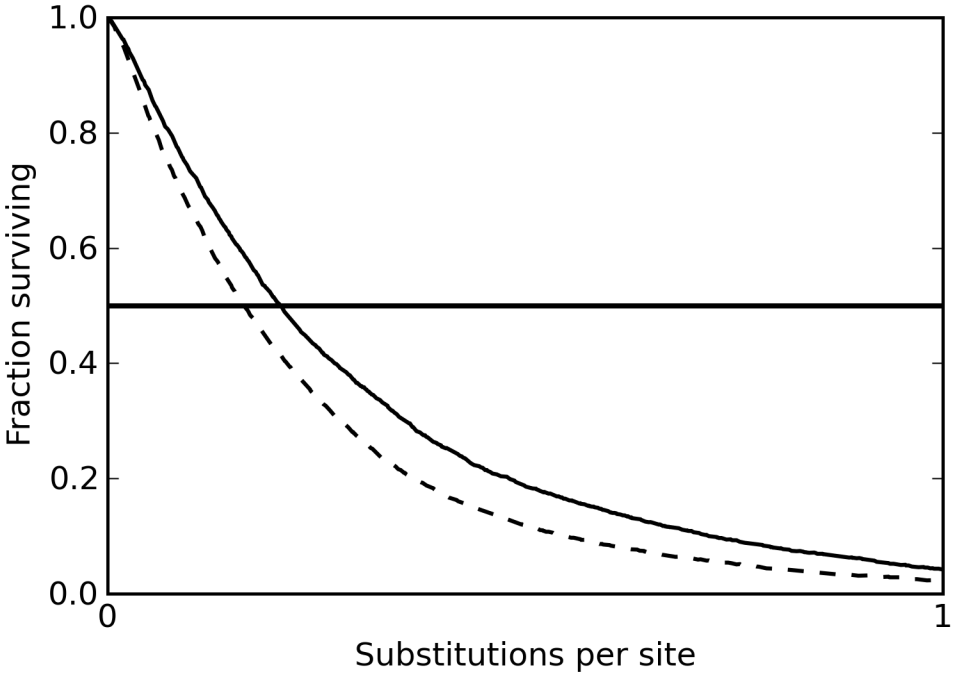


Figure S4. In simulations using the actual *D. melanogaster* substitution and indel patterns, the protective effect of deletions is minimal, as the frequency of multi-site deletions was low relative to single site deletions and point mutations.

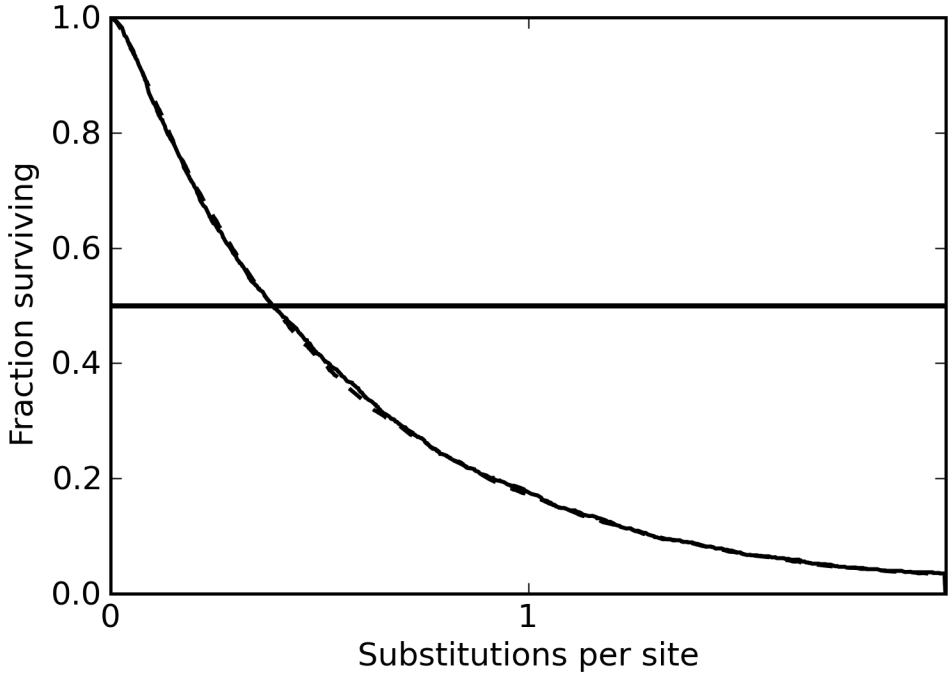


Figure S5. The probability of a KR site containing a BCD site (blue) or vice-versa (red), as described in Figure 1, is plotted as a function of time for rapid (top), normal (middle), and slow (bottom) turnover rates. Rapid turnover was induced by lowering the necessary score thresholds for BCD and KR to 4.5 and 4.6, respectively, and slow turnover induced by raising the necessary score thresholds to 6.5 and 6.6. These simulations have no insertions and deletions.

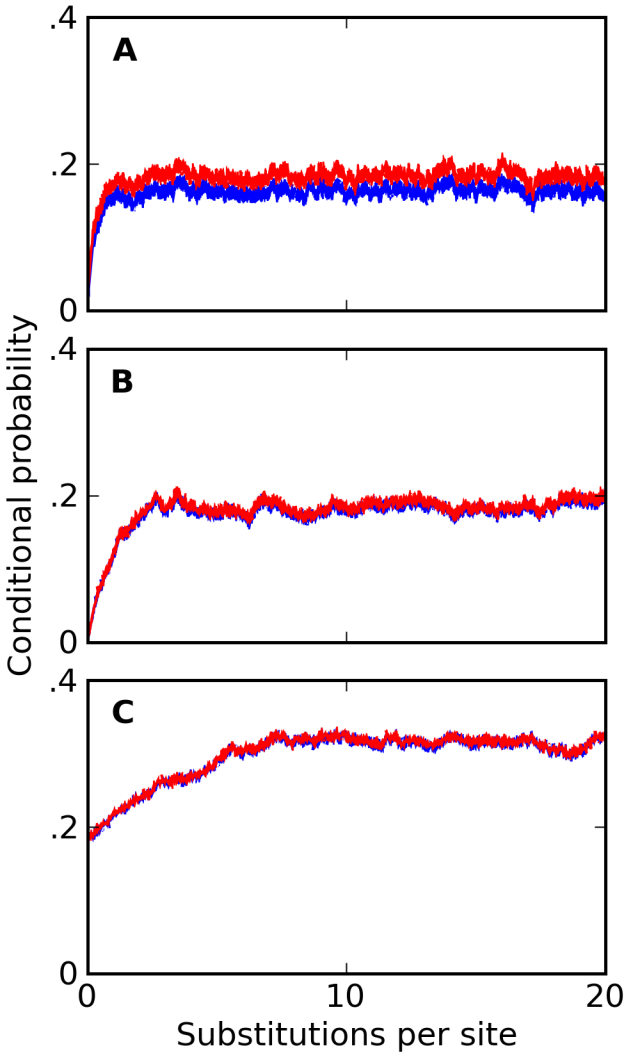


Figure S6. The probability of a KR site containing a BCD site (blue) or vice-versa (red), as described in Figure 1, is plotted as a function of time for rapid (top), normal (middle), and slow (bottom) turnover rates. Rapid turnover was induced by lowering the necessary score thresholds for BCD and KR to 4.5 and 4.6, respectively, and slow turnover induced by raising the necessary score thresholds to 6.5 and 6.6. In these simulations 20% of mutations are indels. The proportion of indels that are deletions is 50% (left), 60% (middle), and 80% (right).

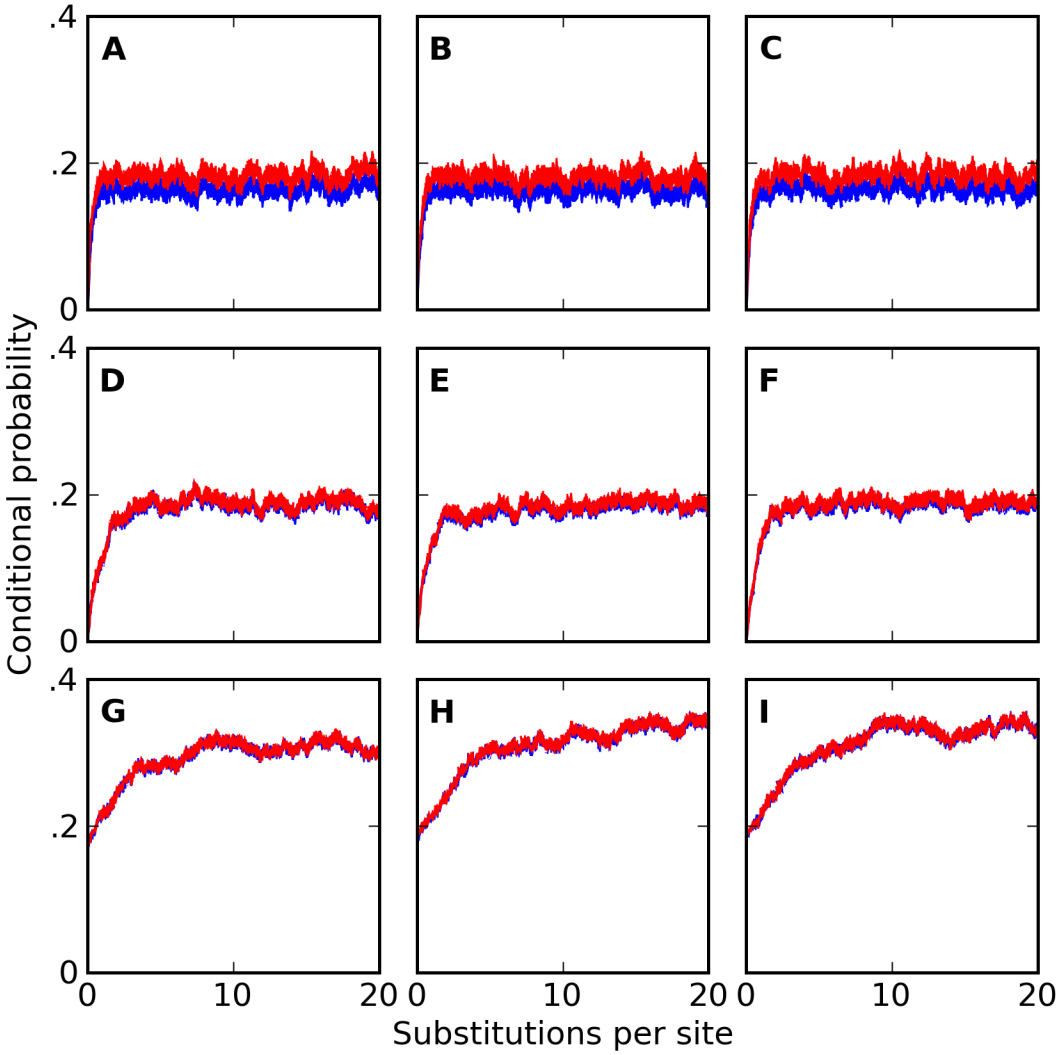
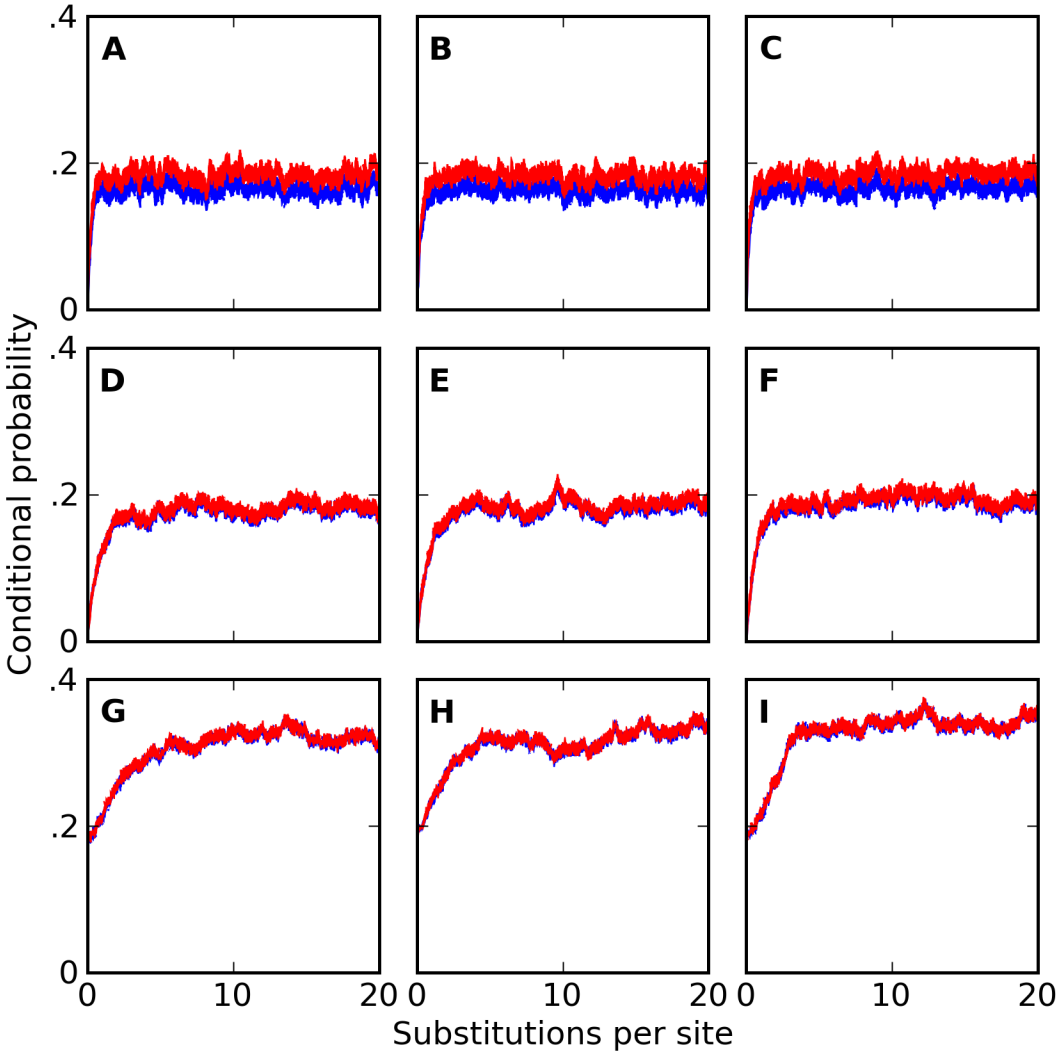


Figure S7. The probability of a KR site containing a BCD site (blue) or vice-versa (red), as described in Figure 1, is plotted as a function of time for rapid (top), normal (middle), and slow (bottom) turnover rates. Rapid turnover was induced by lowering the necessary score thresholds for BCD and KR to 4.5 and 4.6, respectively, and slow turnover induced by raising the necessary score thresholds to 6.5 and 6.6. In these simulations 40% of mutations are indels. The proportion of indels that are deletions is 50% (left), 60% (middle), and 80% (right).



References

1. Pertea M, Salzberg SL (2010). Between a chicken and a grape: estimating the number of human genes. Genome Biol 11(5):206.
2. Levine M, Tjian R (2003). Transcription regulation and animal diversity. Nature 424(6945):147-51.
3. Wray GA (2007). The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206-16.
4. Wray GA (2003). Transcriptional regulation and the evolution of development. Int J Dev Biol 47(7-8):675-84.
5. Roeder RG (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci 21(9):327-35.
6. Nikolov DB, Burley SK (1997). RNA polymerase II transcription initiation: a structural view. Proc Natl Acad Sci U S A 94(1):15-22.
7. Lee TI, Young RA (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34:77-137.
8. Smale ST, Kadonaga JT (2003). The RNA polymerase II core promoter. Annu Rev Biochem 72:449-79.
9. Näär AM, Lemon BD, Tjian R (2001). Transcriptional coactivator complexes. Annu Rev Biochem 70:475-501.
10. Goodrich JA, Tjian R (2010). Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation. Nat Rev Genet 11(8):549-58.
11. Arnosti DN, Kulkarni MM (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem 94(5):890-8.
12. van Nimwegen E (2003). Scaling laws in the functional content of genomes. Trends Genet 19(9):479-84.
13. Shore D, Nasmyth K (1987). Purification and cloning of a DNA binding protein from yeast that binds to both silencer and activator elements. Cell 51(5):721-32.
14. Kadonaga JT (2004). Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116(2):247-57.
15. Ptashne M, Gann A (1997). Transcriptional activation by recruitment. Nature 386(6625):569-77.
16. Malik S, Roeder RG (2010). The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. Nat Rev Genet 11(11):761-72.
17. Levine M (2010). Transcriptional enhancers in animal development and evolution. Curr Biol 20(17):R754-63.
18. Cai HN, Arnosti DN, Levine M (1996). Long-range repression in the Drosophila embryo. Proc Natl Acad Sci U S A 93(18):9309-14.
19. Thanos D, Maniatis T (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83(7):1091-100.
20. Stanojevic D, Small S, Levine M (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. Science 254(5036):1385-7.
21. Small S, Blair A, Levine M (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J 11(11):4047-57.

22. Arnosti DN, Barolo S, Levine M, Small S (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development 122(1):205-14.
23. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res 38(Database
24. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS (2003). Global analysis of protein expression in yeast. Nature 425(6959):737-41.
25. Stormo GD (2000). DNA binding sites: representation and discovery. Bioinformatics 16(1):16-23.
26. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 10(9):2997-3011.
27. Staden R (1989). Methods for calculating the probabilities of finding patterns in sequences. Comput Appl Biosci 5(2):89-96.
28. Claverie JM, Audic S (1996). The statistical significance of nucleotide position-weight matrix matches. Comput Appl Biosci 12(5):431-9.
29. Tuerk C, Gold L (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249(4968):505-10.
30. Galas DJ, Schmitz A (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5(9):3157-70.
31. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res 19(4):556-66.
32. Siddharthan R (2010). Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. PLoS One 5(3):e9722.
33. Man TK, Stormo GD (2001). Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res 29(12):2471-8.
34. Bulyk ML, Johnson PL, Church GM (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 30(5):1255-61.
35. Zhou Q, Liu JS (2004). Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 20(6):909-16.
36. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99(2):757-62.
37. Erives A, Levine M (2004). Coordinate enhancers share common organizational features in the Drosophila genome. Proc Natl Acad Sci U S A 101(11):3851-6.
38. Kim J, He X, Sinha S (2009). Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet 5(1):e1000330.
39. Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA (2003). Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription

regulatory information. Nucleic Acids Res 31(20):6016-26.
40. Papatsenko D, Goltsev Y, Levine M (2009). Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res 37(17):5665-77.
41. Aparicio O, Geisberg JV, Struhl K (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol Chapter.
42. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000). Genome-wide location and function of DNA binding proteins. Science 290(5500):2306-9.
43. Johnson DS, Mortazavi A, Myers RM, Wold B (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830):1497-502.
44. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004). Transcriptional regulatory code of a eukaryotic genome. Nature 431(7004):99-104.
45. Wilbanks EG, Facciotti MT (2010). Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 5(7):e11471.
46. Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, Iyer VN, Hechmer A, Simirenko L, Stapleton M, Luengo Hendriks CL, Chu HC, Ogawa N, Inwood W, Sementchenko V, Beaton A, Weiszmann R, Celniker SE, Knowles DW, Gingeras T, Speed TP, Eisen MB, Biggin MD (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol 6(2):e27.
47. Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Bristow J, Ren B, Black BL, Rubin EM, Visel A, Pennacchio LA (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet 42(9):806-10.
48. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457(7231):854-8.
49. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol 5(9):R61.
50. 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.
51. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol 5(12):R98.
52. Siddharthan R (2008). PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. PLoS Comput Biol 4(8):e1000156.
53. Sinha S (2007). PhyME: a software tool for finding motifs in sets of orthologous sequences. Methods Mol Biol 395:309-18.
54. Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA (2008). Ultraconservation identifies a small subset of

extremely constrained developmental enhancers. Nat Genet 40(2):158-60.
55. Peterson BK, Hare EE, Iyer VN, Storage S, Conner L, Papaj DR, Kurashima R, Jang E, Eisen MB (2009). Big genomes facilitate the comparative identification of regulatory elements. PLoS One 4(3):e4688.
56. Ludwig MZ, Patel NH, Kreitman M (1998). Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 125(5):949-58.
57. Ludwig MZ, Kreitman M (1995). Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. Mol Biol Evol 12(6):1002-11.
58. MacArthur S, Brookfield JF (2004). Expected rates and modes of evolution of enhancer sequences. Mol Biol Evol 21(6):1064-73.
59. Dermitzakis ET, Clark AG (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19(7):1114-21.
60. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet 4(6):e1000106.
61. Bulyk ML (2003). Computational prediction of transcription-factor binding site locations. Genome Biol 5(1):201.
62. Sharon E, Lubliner S, Segal E (2008). A feature-based approach to modeling protein-DNA interactions. PLoS Comput Biol 4(8):e1000154.
63. von Hippel PH, Berg OG (1989). Facilitated target location in biological systems. J Biol Chem 264(2):675-8.
64. Heumann JM, Lapedes AS, Stormo GD (1994). Neural networks for determining protein specificity and multiple alignment of binding sites. Proc Int Conf Intell Syst Mol Biol 2:188-94.
65. Segal E, Barash Y, Simon I, Friedman N, Koller D (2002). From promoter sequence to expression: a probabilistic framework. RECOMB 2002: 263-272.
66. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. BMC Bioinformatics 7:113.
67. Tanay A (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res 16(8):962-72.
68. Berg OG, von Hippel PH (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol 193(4):723-50.
69. Hasegawa M, Kishino H, Yano T (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22(2):160-74.
70. Halpern AL, Bruno WJ (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15(7):910-7.
71. Boros J, Lim FL, Darieva Z, Pic-Taylor A, Harman R, Morgan BA, Sharrocks AD (2003). Molecular determinants of the cell-cycle regulated Mcm1p-Fkh2p transcription factor complex. Nucleic Acids Res 31(9):2279-88.
72. Mao C, Carlson NG, Little JW (1994). Cooperative DNA-protein interactions. Effects of

changing the spacing between adjacent binding sites. J Mol Biol 235(2):532-44.
73. Ioshikhes I, Trifonov EN, Zhang MQ (1999). Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc Natl Acad Sci U S A 96(6):2891-5.
74. Lipman DJ, Pearson WR (1985). Rapid and sensitive protein similarity searches. Science 227(4693):1435-41.
75. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Res 26(1):73-9.
76. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, NISC Comparative Sequencing Program, Green ED, Sidow A, Batzoglou S (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13(4):721-31.
77. Hertz GZ, Stormo GD (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15(7-8):563-77.
78. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M (2007). Divergence of transcription factor binding sites across related yeast species. Science 317(5839):815-9.
79. Chin CS, Chuang JH, Li H (2005). Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. Genome Res 15(2):205-13.
80. Maerkl SJ, Quake SR (2007). A systems approach to measuring the binding energy landscapes of transcription factors. Science 315(5809):233-7.
81. Gertz J, Siggia ED, Cohen BA (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 457(7226):215-8.
82. Lam FH, Steger DJ, O'Shea EK (2008). Chromatin decouples promoter threshold from dynamic range. Nature 453(7192):246-50.
83. Prado JL, Limãos EA, Roblero J, Freitas JO, Prado ES, Paiva AC (1975). Recovery and conversion of kinins in exsanguinated rat preparations. Naunyn Schmiedebergs Arch Pharmacol 290(2-3):191-205.
84. Piña B, Fernández-Larrea J, García-Reyero N, Idrissi FZ (2003). The different (sur)faces of Rap1p. Mol Genet Genomics 268(6):791-8.
85. Idrissi FZ, Garcia-Reyero N, Fernandez-Larrea JB, Piña B (2001). Alternative mechanisms of transcriptional activation by Rap1p. J Biol Chem 276(28):26090-8.
86. Hartley PD, Madhani HD (2009). Mechanisms that specify promoter nucleosome location and identity. Cell 137(3):445-58.
87. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. Nature 446(7135):572-6.
88. Blaiseau PL, Thomas D (1998). Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. EMBO J 17(21):6327-36.
89. Tanaka M (1996). Modulation of promoter occupancy by cooperative DNA binding and activation-domain function is a major determinant of transcriptional regulation by activators in vivo. Proc Natl Acad Sci U S A 93(9):4311-5.
90. Kim JH, Polish J, Johnston M (2003). Specificity and regulation of DNA binding by the yeast

glucose transporter gene repressor Rgt1. Mol Cell Biol 23(15):5208-16.
91. Kim JH (2009). DNA-binding properties of the yeast Rgt1 repressor. Biochimie 91(2):300-3.
92. Felenbok B, Flipphi M, Nikolaev I (2001). Ethanol catabolism in Aspergillus nidulans: a model system for studying gene regulation. Prog Nucleic Acid Res Mol Biol 69:149-204.
93. Miller JA, Widom J (2003). Collaborative competition mechanism for gene activation in vivo. Mol Cell Biol 23(5):1623-32.
94. Mirny LA (2009). Nucleosome-mediated cooperativity between transcription factors. Nat Precedings 2009; http://hdl.handle.net/10101/npre.2009.2796.1.
95. Elemento O, Slonim N, Tavazoie S (2007). A universal framework for regulatory element discovery across all genomes and data types. Mol Cell 28(2):337-50.
96. Westholm JO, Xu F, Ronne H, Komorowski J (2008). Genome-scale study of the importance of binding site context for transcription factor binding and gene regulation. BMC Bioinformatics 9:484.
97. Nguyen DH, D'haeseleer P (2006). Deciphering principles of transcription regulation in eukaryotic genomes. Mol Syst Biol 2:2006.0012.
98. Lin Z, Wu WS, Liang H, Woo Y, Li WH (2010). The spatial distribution of cis regulatory elements in yeast promoters and its implications for transcriptional regulation. BMC Genomics 11:581.
99. Wunderlich Z, Mirny LA (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet 25(10):434-40.
100. Danchin A (2000). A brief history of genome research and bioinformatics in France. Bioinformatics 16(1):65-75.
101. Martínez-Pastor MT, Marchler G, Schüller C, Marchler-Bauer A, Ruis H, Estruch F (1996). The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). EMBO J 15(9):2227-35.
102. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11(12):4241-57.
103. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, Jennings EG, Lee TI, True HL, Lander ES, Young RA (2001). Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell 12(2):323-37.
104. Hu Z, Killion PJ, Iyer VR (2007). Genetic reconstruction of a functional transcriptional regulatory network. Nat Genet 39(5):683-7.
105. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320(5881):1344-9.
106. Nibu Y, Senger K, Levine M (2003) CtBP-independent repression in the Drosophila embryo. Mol Cell Biol 23: 3990-3999.
107. Kulkarni MM, Arnosti DN (2005) cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. Mol Cell Biol 25: 3411-3420.
108. Lebrecht D, Foehr M, Smith E, Lopes FJ, Vanario-Alonso CE, et al. (2005) Bicoid cooperative DNA binding is critical for embryonic patterning in Drosophila. Proc Natl Acad Sci U S A 102: 13176-13181.

109. Kulkarni MM, Arnosti DN (2003) Information display by transcriptional enhancers. Development 130: 6569-6575.
110. Merika M, Thanos D (2001) Enhanceosomes. Curr Opin Genet Dev 11: 205-208.
111. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104 Suppl 1: 8597-8604.
112. Tanay A, Siggia ED (2008) Sequence context affects the rate of short insertions and deletions in flies and primates. Genome Biol 9: R37.
113. He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. PLoS Comput Biol 5: e1000299.
114. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2: e130.
115. Driever W, Nusslein-Volhard C (1988) A gradient ofbicoid protein in Drosophila embryos. Cell 54: 83-93.
116. Stanojevic D, Hoey T, Levine M (1989) Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in Drosophila. Nature 341: 331-335.
117. Treisman J, Desplan C (1989) The products of the Drosophila gap genes hunchback and Kruppel bind to the hunchback promoters. Nature 341: 335-337.
118. Rivera-Pomar R, Jackle H (1996) From gradients to stripes in Drosophila embryogenesis: filling in the gaps. Trends Genet 12: 478-483.
119. Small S, Kraut R, Hoey T, Warrior R, Levine M (1991) Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev 5: 827-839.
120. Petrov DA (2002) DNA loss and evolution of genome size in Drosophila. Genetica 115: 81-91.
121. Bergman CM, Carlson JW, Celniker SE (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747-1749.
122. Liang HL, Nien CY, Liu HY, Metzstein MM, Kirov N, et al. (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature 456: 400-403.
123. Rastegar S, Hess I, Dickmeis T, Nicod JC, Ertzer R, et al. (2008) The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. Dev Biol 318: 366-377.
124. Won KJ, Sandelin A, Marstrand TT, Krogh A (2008) Modeling promoter grammars with evolving hidden Markov models. Bioinformatics 24: 1669-1675.
125. Neafsey DE, Palumbi SR (2003) Genome size evolution in pufferfish: a comparative analysis of diodontid and tetraodontid pufferfish genomes. Genome Res 13: 821-830.
126. Graur D, Shuali Y, Li WH (1989) Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol 28: 279-285.
127. Petrov DA, Hartl DL (1998) High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol Biol Evol 15: 293-302.
128. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) Evidence for DNA loss as a determinant of genome size. Science 287: 1060-1062.
129. Robertson HM (2000) The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res 10: 192-203.

130. Bensasson D, Petrov DA, Zhang DX, Hartl DL, Hewitt GM (2001) Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol Biol Evol 18: 246-253.
131. Brown CT, Xie Y, Davidson EH, Cameron RA (2005) Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison. BMC Bioinformatics 6: 70.
132. Down TA, Bergman CM, Su J, Hubbard TJ (2007) Large-scale discovery of promoter motifs in Drosophila melanogaster. PLoS Comput Biol 3: e7.
133. Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, et al. (2008) A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Res 36: 2547-2560.
134. De Renzis S, Elemento O, Tavazoie S, Wieschaus EF (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol 5: e117.

Appendix A

A population-genetic model behaves similarly to the threshold model

Whole-population simulations can provide a precise view of how selection impacts sequence change, although they can be computationally expensive. We used such simulations to explore the evolutionary properties of enhancer elements under selection to maintain a given binding site composition, but we were unable to find differences between the results of these simulations and those presented in the main text.

Like those simulations, these required an enhancer to have at least a certain number of binding sites for a specified set of factors. In this case, each enhancer was required to have five Bicoid sites and five Kruppel sites. Each generation had a mutation step and a selection step. However, instead of being rejected outright, mutations to enhancers that brought the number of sites below these requirements were given a selective penalty defined as the number of missing sites multiplied by a penalty factor *s*. In the selection step, alleles were resampled according to their selective penalties. This method gives even deleterious alleles missing binding sites some chance of being fixed in the population.

We ran our simulations of a population of 10,000 enhancers for five values of the selective penalty factor *s*. While the lowest value of *s* we tested did not appreciably increase the number of binding sites above that expected by chance (fig. 1), the intermediate value of $s = .0001$ provided an interesting case. Compared to simulations run under more stringent selective penalties, not only was the rate of loss per site greater, but also, less intuitively, the rate of site gain from neutral sequence was greater (fig. 3). As the rate of generation of alleles containing new sites must remain roughly the same between these simulations, we reasoned that the probability of fixation of alleles containing new sites must be higher. Indeed, when a more stringent selective penalty was used, virtually no new alleles arose with a selective advantage over the major allele, as all alleles at appreciable frequency already have a sufficient number of sites (data not shown). At this lesser penalty, where the major allele typically has an insufficient number of sites, alleles containing new sites can carry a selective advantage. While this phenomenon increased the rate of turnover, it had no effect on the increased enrichment of overlapping binding sites.

At values of *s* sufficiently large to maintain the required number of sites (s=.01, s=.001, fig. 1), the rate of turnover, spatial distribution of sites, and enrichment of overlapping binding sites were indistinguishable from those properties as observed in the simpler threshold-based model (figs. 2-4), allowing us to take advantage of the greater computational tractability of the threshold model in the main text.

Methods

Each simulation was started from a population containing 10,000 identical 1,000 base pair enhancers with five Bicoid sites and five Kruppel sites meeting the score cutoffs described in the main text. For each generation's mutation step, we used the mutation rate described in [1] to arrive at a mutation rate of 7.56*10^-6 mutations per enhancer per generation. Assuming that no single enhancer would mutate twice in a single generation, we sampled from a poisson to determine the number of new alleles to create. Each new allele was generated from a randomly selected enhancer by sampling from the mutation distributions described in the main text, with a 20% chance of creating an indel and a 60% chance of creating a deletion if an indel was chosen.

In each generation's selection step, we sampled 10,000 new alleles from a multinomial distribution. The parameters of this distribution were determined by creating a weight for each allele defined as that allele's current-generation count multiplied by its selective penalty, these weights then being normalized to sum to one.

For each tested penalty factor, seventy-five replicates were tested to 1.32275 billion generations, the time necessary for 10,000 neutral mutations to reach fixation. Every 5.291 million generations, or forty fixed neutral mutations, the major allele was recorded, each time being surveyed for binding sites for Bicoid and Kruppel. This chain of alleles was aligned using FSA [2], allowing us to determine the amount of lost and gained sites per substitution. Although we expect alignment errors at this distance to be minimal, to best compare these results with the model of the main text, we output the sequence of those simulations every forty mutation-selection iterations and created a similar chain of aligned sequences from which gain and loss statistics were derived.

To determine the count and spacing distribution of sites and the enrichment of overlapped sites, major alleles were sampled at 2.5, 5.0, 7.5, and 10.0 substitutions per site. 95% confidence intervals for number of sites, turnover rate, site spacing, and overlap effect were calculated by resampling the data over 1,000 bootstrap replicates. Comparisons between spacer length distributions were made between the cutoff model and the population genetic model using Pearson's chi-squared test.

Figure 1. The average number of sites depends on the selective penalty for missing sites. *s* greater than or equal to .001 is necessary to maintain the required number of sites (10). At these values, the average number of sites is not distinguishable from the average number given by the cutoff model.

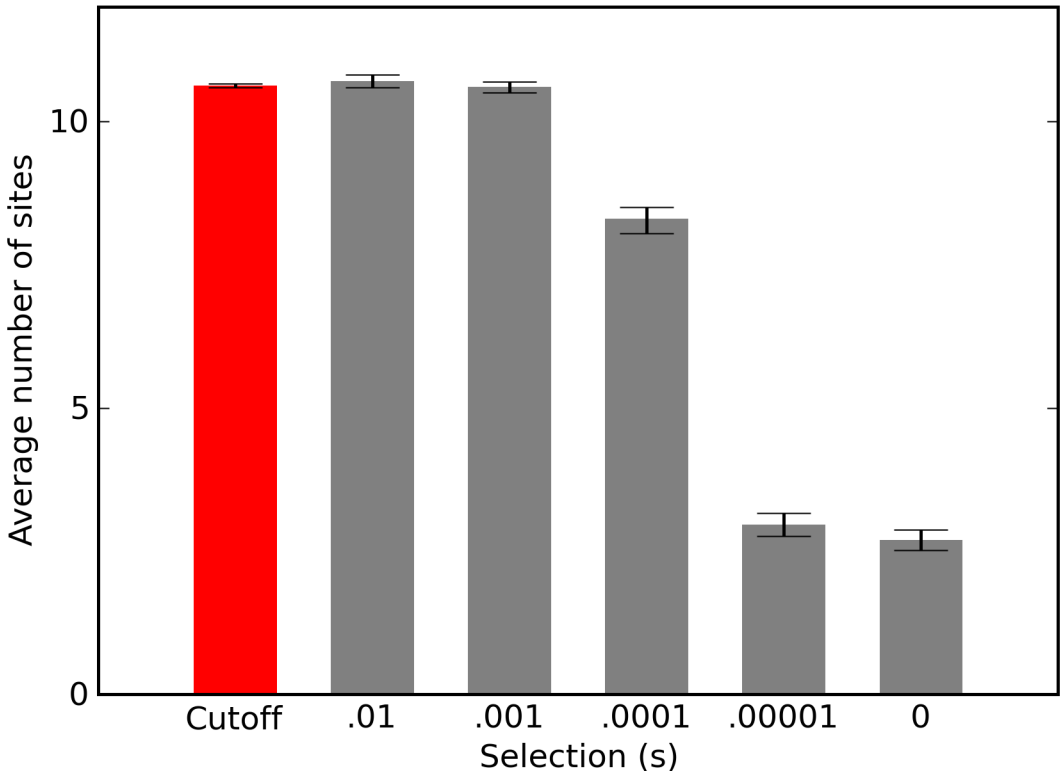


Figure 2. A population-genetic model does not affect the spatial distribution of binding sites. Spacer elements between binding sites were divided into five distance bins, and the fraction of all spacers in each bin was plotted for the cutoff model and five population genetic models with different values of *s* (legend). An * denotes a significant difference from the cutoff model at α = .05.

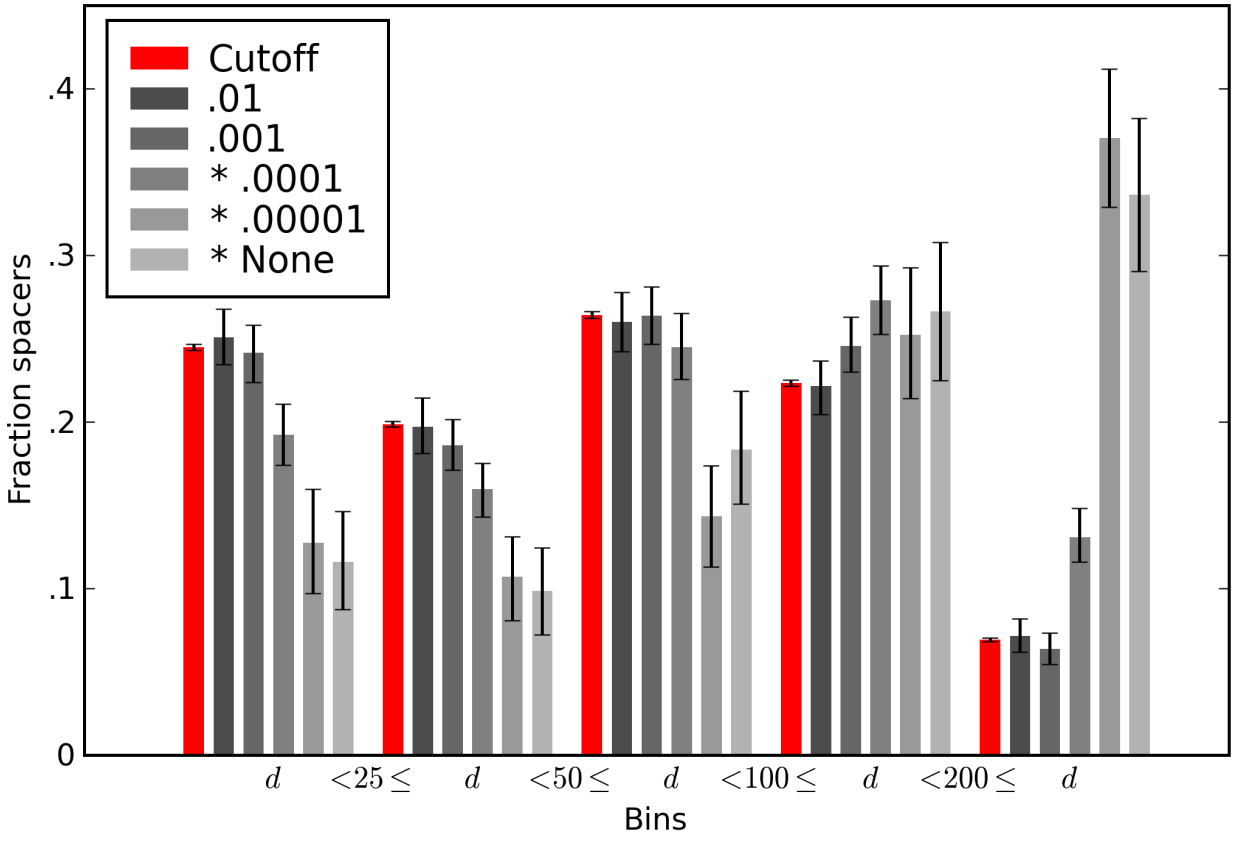


Figure 3. Rates of turnover are dependent upon selective penalty. For each value of *s*, rates of binding site loss (left) and gain (right) per neutral substitution were plotted. At values of *s* sufficient to maintain binding site composition, the rate of turnover is not significantly different from that observed under the cutoff model.

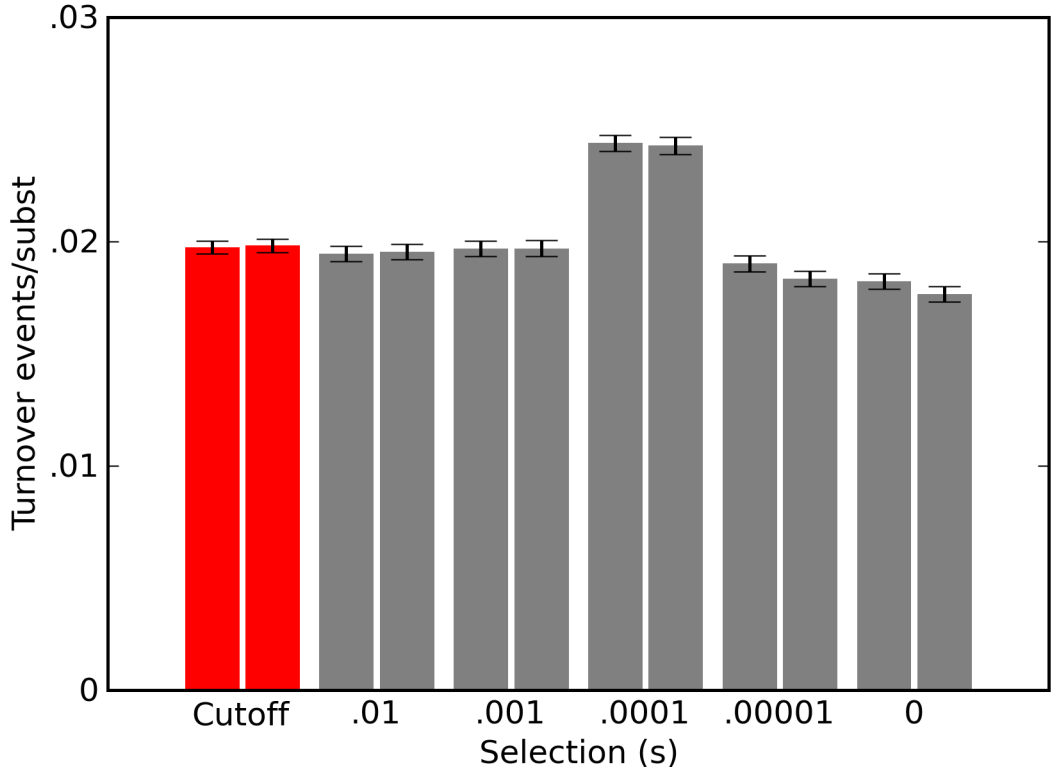
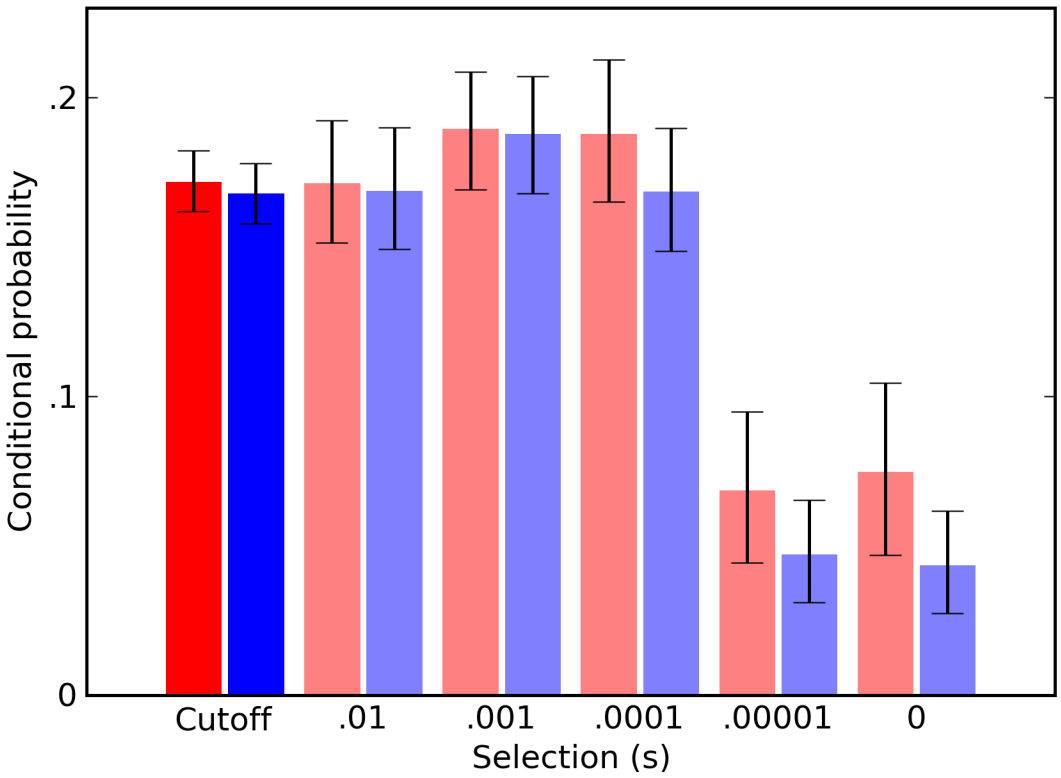


Figure 4. Overlapping binding sites are enriched in a population genetic model. The post-simulation probability of observing a Kruppel site conditioned on seeing a Bicoid site (blue) and a Bicoid site conditioned on seeing a Kruppel site (red) is similar to that observed in the cutoff model when the selective penalty is sufficient to markedly increase the number of binding sites (s greater than or equal to .0001).



References

1. Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL et al (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. Nature 445: 82-85.

2. Bradley RK, Roberts A, Smoot M, Juvekar S, Do J et al (2009). Fast statistical alignment. PLoS Comput Biol 5: e1000392**.**

Appendix B

An evolutionary model of overlapping sites predicts a reduced nucleotide substitution rate

As an alternative test of overlapping versus singleton site conservation, we extended a base-specific model of binding site evolution to the case of overlapping binding sites. To this end, we made the following two assumptions: first, we assumed that if two binding sites overlap but do not have a biologically meaningful interaction, then mutations in the overlapped region should behave as simultaneous mutations in two non-overlapping binding sites. Second, we assumed that selection pressure is additive. We used the framework developed by Halpern and Bruno [1], which has been shown to accurately model the evolution of individual transcription factor binding sites [2-5], to calculate the selection coefficients of every possible mutation as made to each individual site, and then we combined these coefficients to arrive at the coefficient for every possible mutation in the overlapped site. We then used these coefficients then determine position-specific rates of substitution according to classic relationships developed by Kimura [6].

We used the above described model to derive rates of substitution relative to the neutral rate for an overlapped Bicoid/Krüppel site. In positions where the two factors share a nucleotide preference, the substitution rate is strongly lowered (fig. 1), and the substitution in the overlapped region had substitution rates as low as 3.4% of the predicted single-factor rate (fig. 2). This result was consistent with longer half-lives of overlapping sites in our simulations.

Methods

According to [1], for base frequencies pi and neutral substitution rates p, 2Ns for any given mutation is:

$$2\mathrm{Ns} = \log\left(\frac{\pi_b p_{ba}}{\pi_a p_{ab}}\right) \quad (1)$$

According to the equations of Kimura [24], the fixation probability of such an allele is:

$$f_{ab} \approx \frac{2\mathrm{s}}{1 - e^{-2\mathrm{Ns}}} \quad (2)$$

Assuming additive selection, the rates of substitution for an overlapped region of two sites are:

$$f_{ab} \propto \frac{\log(HB1) + \log(HB2)}{1 - e^{-\log(HB1) - \log(HB2)}} \quad (3)$$

where

$$HB = \frac{\pi_b p_{ba}}{\pi_a p_{ab}} \quad (4)$$

Figures

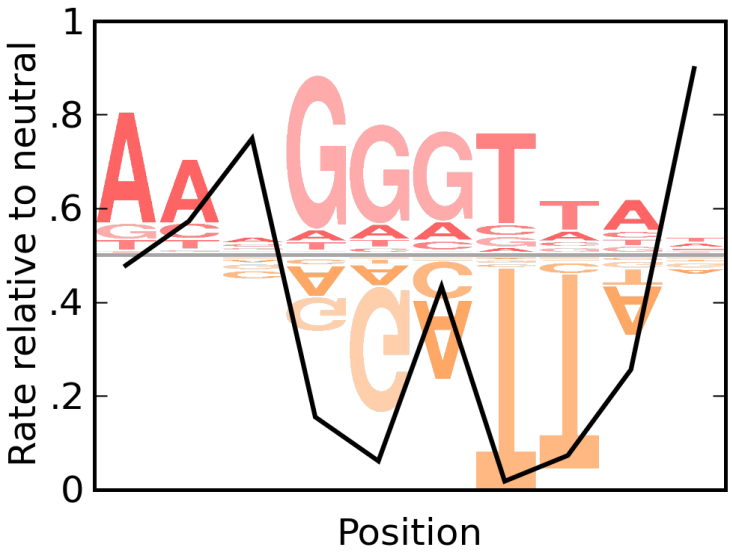


Figure 1. .Aligned nucleotide preferences lower substitution rate.
Predicted rates of evolution of an overlapping Bicoid/Krüppel binding site, with the sequence logos of Krüppel (top, red) and Bicoid (bottom, orange) in the background. The rate (black) is taken relative to the expected neutral rate at that position.

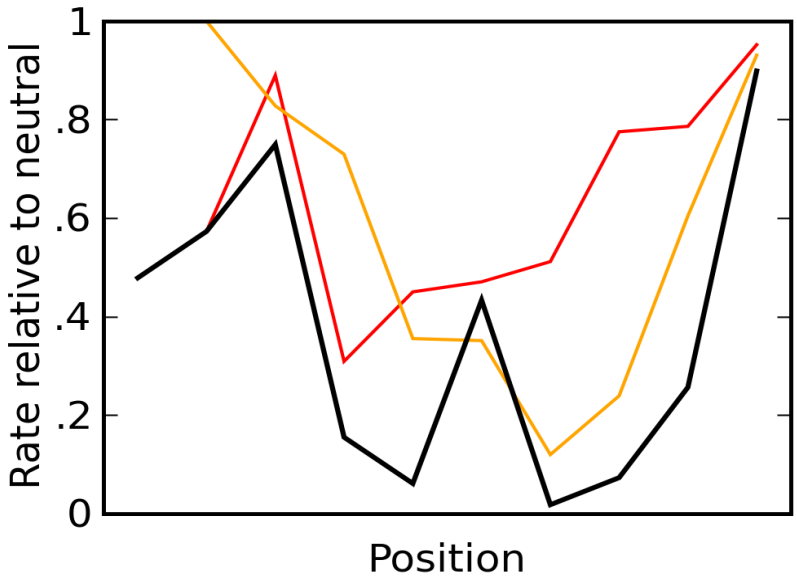


Figure 2. Site overlap strongly reduces substitution rate
Predicted rates of evolution of a Krüppel site (red), a Bicoid site (orange) and an overlapping Bicoid/Krüppel site (black).

References

1. Halpern A, Bruno W (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15(7): 910-917.
2. He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. PLoS Comput Biol 5(3): e1000299. doi:10.1371/journal.pcbi.1000299
3. Doniger S, Fay J (2007) Frequent Gain and Loss of Functional Transcription Factor Binding Sites. PLoS Comput Biol 3(5): e99. doi:10.1371/journal.pcbi.0030099
4. Moses A, Pollard D, Nix D, Iyer V, Li X et al (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2(10): e130. doi:10.1371/journal.pcbi.0020130
5. Moses A, Chiang D, Kellis M, Lander E, Eisen M (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
6. Kimura M (1962) On the probability of fixation of mutant genes in a population. Genetics 47: 713-719.

Appendix C

The frequency and size of insertions and deletions affect site clustering

Several recent works have proposed the importance of overlapping and locally clustered sites within *Drosophila* [1-4], and so we parameterized our neutral mutation model to approximate the patterns of point mutations, insertions, and deletions found within that clade. Although we expected that the enrichment and increased conservation of overlapping sites should largely depend on the sequence specificities of the two transcription factors involved, it was clear that the spatial arrangement of binding sites should be informed by the parameters of the neutral mutation model, in particular the indel rate and the bias, if any, of indels towards deletions. In order to generalize the quantitative conclusions of the main text beyond *Drosophila*, we trained a predictive model of binding site arrangement on simulated enhancers evolved under a wide range of these parameters. We were able to accurately predict the divergence of spatial distributions of binding sites within these enhancers from that within indel-free enhancers using only the frequencies of deletions and insertions (fig. 1).

To highlight this effect, we chose to investigate in greater detail several species whose patterns of indels have been characterized. We simulated the evolution of enhancers according to parameters derived from *C. elegans*, which has a high rate of DNA loss, mammals, and two species of grasshoppers, one of which has a particularly low rate of DNA loss [5]. In addition, we investigated the effect of another set of indel parameters for *D. melanogaster* different from that used in the main text. In these simulations, not only the frequencies of insertions and deletions but also their average size were incorporated. We found, again, that while in each case the distribution of spacer elements was skewed, the magnitude of this skew varied widely depending on the choice of parameters: *C. elegans* and the alternate parameterization of *D. melanogaster* showed a substantially stronger enrichment of locally clustered sites than observed in the main text, while other species showed a weaker enrichment (fig. 2).

Methods

To train the model, we performed 750 simulations for each pair of indel mutation parameters shown in fig. 1. Each simulation evolved a 1,000bp enhancer containing 10 Kruppel sites (score > 5.6) for 30,000 mutation-selection rounds (simulation details are available in the main text). In these simulations, to better allow interpretation an even deletion bias, insertions and deletions were of equal average length, their lengths both being drawn from the distribution of deletion lengths described in the main text. Spacers between non-overlapping binding sites were binned according to cutoffs 25, 50, 75, 100, 150, 200, 250, 300, and 500 base pairs. The Kullback-Leibler divergence was calculated between each of these samples and a 'reference' 5,000-replicate set derived from simulations without indels. We fit a linear model relating the KL divergence to the insertion and deletion frequencies using R (ID is indel frequency, DB is deletion bias, and KL is KL divergence):

```
KL = -.147*ID + .101*DB + .454*ID*DB - .044
```

The highlighted species used indel rate, average indel sizes, and deletion bias data from [5]. The indel sizes were transformed into geometric distributions from which lengths were sampled during the simulations.

Figures

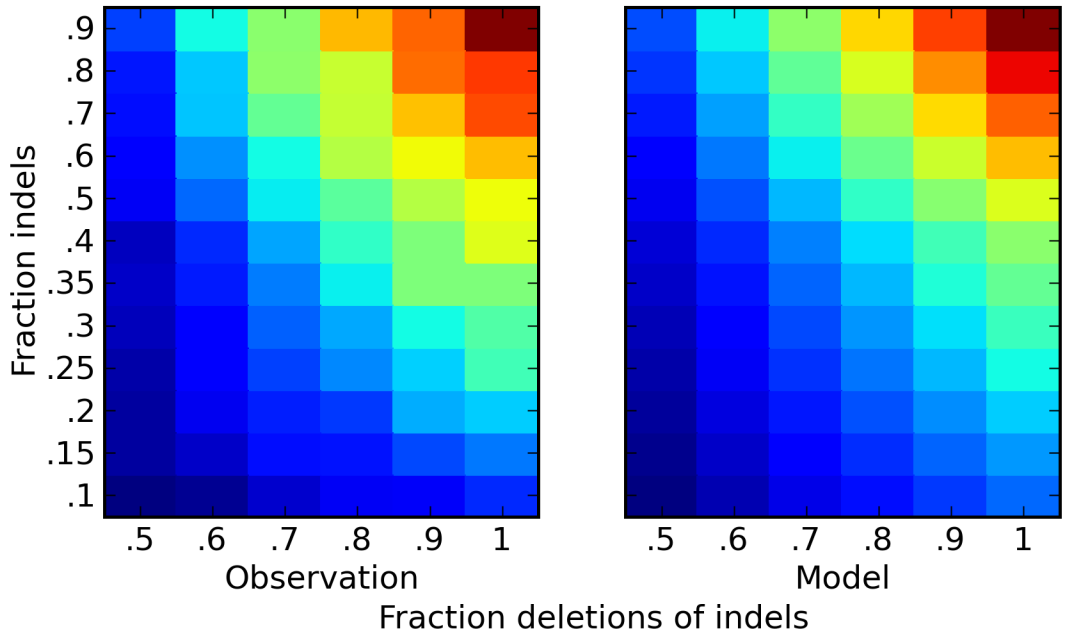


Figure 1. Deviation in spacing distribution is predictable by a linear model. Kullback-Leibler divergence is plotted by color from a minimum of .0145 (deep blue) to a maximum of .334 (deep red).

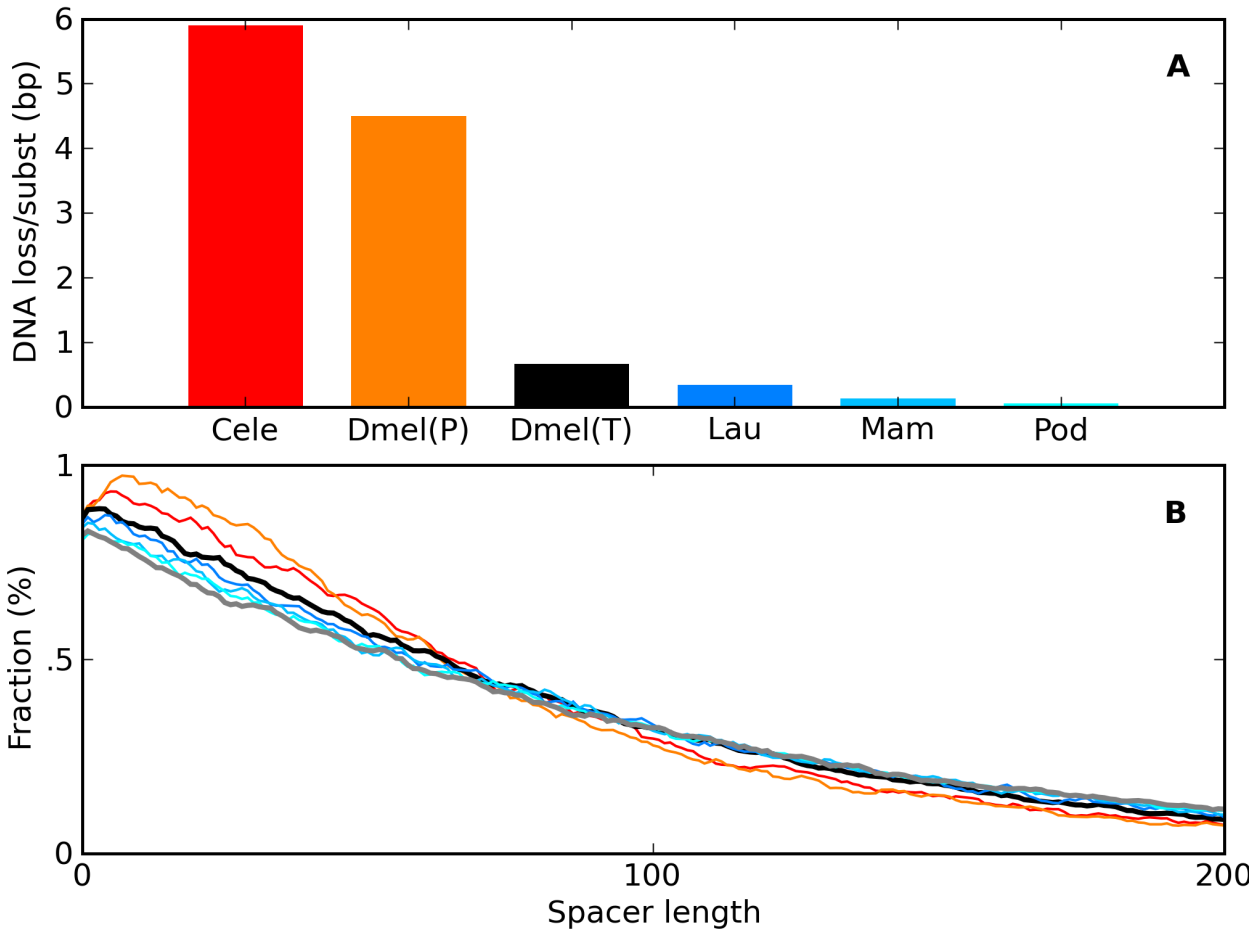


Figure 2. Spacing distribution skew is different in different organisms and depends on the rate of DNA loss. A. DNA loss, as calculated and described in [5]. Cele is *C. elegans*, Lau and Pod refer to *Laupala* and *Podisma* grasshoppers, Mam is mammals, and Dmel(P) and Dmel(T) refer to *D. melanogaster* as described in [5] and [6], respectively. The main text used the parameters of Dmel(T). B. Spacer length distribution averaged over five base pair windows. Colors are as in (a), except for the gray line, which refers to simulations without indels.

References

1. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet 4: e1000106.
2. Kim J, He X, Sinha S (2009) Evolution of regulatory sequences in 12 Drosophila species. PLoS Genet 5: e1000330.
3. Makeev VJ, Lifanov AP, Nazina AG, Papatsenko DA (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res 31: 6016-6026.

4. Papatsenko D, Goltsev Y, Levine M (2009) Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res.
5. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) Evidence for DNA loss as a determinant of genome size. Science 287: 1060-1062.
6. Tanay A, Siggia ED (2008) Sequence context affects the rate of short insertions and deletions in flies and primates. Genome Biol 9: R37.

Appendix D

Testing synthetic enhancers in transgenic embryos

Introduction

In the main text, we called into question evidence that spatially restricted 'grammars' of linked binding sites are necessary for enhancer function, showing that much of this evidence arises as a byproduct of selection on binding site composition alone. The follow-up question is intuitive: do synthetic enhancers that are spatially scrambled but have conserved binding site composition retain the expression activity of the original?

Not all of the evidence supporting the grammar hypothesis has come from the sequence and evolutionary analyses we discuss in chapter four. Indeed, substantial effort has been made towards, on a genetic level, testing the effect of perturbed spatial arrangements of binding sites within enhancers. These efforts came to sometimes contradictory conclusions: as discussed in detail in the introduction to the main text, some enhancers appear to serve as scaffolds for the precise assembly of multiprotein complexes, making them sensitive to any kind of disruption, while others, including the enhancer we investigate in detail in chapter four, are robust to even radical-seeming changes. Given this diversity of results, computational and evolutionary analysis of whole genomes held great promise for understanding what mechanism governed the operation of the typical enhancer. This analysis appeared to show qualified support for the latter hypothesis: while the great diversity of binding site arrangements exhibited by different species suggested that their arrangement was flexible, certain classes of binding sites, those that were clustered and/or overlapping, appeared to be both enriched and conserved. Hence, enhancers appeared to operate with global flexibility but local constraint. As we showed that this computational evidence for local constraint is plausibly artifactual, here we return to an experimental approach to determine whether local spatial interactions truly constraint the function of enhancer sequences.

Here we have two principle advantages over previous work. First, and chiefly, wholesale synthesis of enhancer-scale sequences has become economical, allowing us much greater flexibility in probing local spatial constraints. Second, our better understanding of how transcription factors interact with DNA, e.g. the importance of weak sites and our nuanced descriptions of transcription factor affinities, should in principle allow us to better recognize and conserve binding site composition. The artificial enhancers described here will shed light on the importance of local spatial constraint in Drosophila developmental enhancers.

Results & Discussion

Eve stripe two binding site composition varies across the 12 fly genomes

The first step in generating spatially scrambled fly sequences is the assembly of an appropriate training set of enhancers that we can use to formally describe binding site composition. Using several sequences, as opposed to using only the D. melanogaster sequence, has two principal advantages: first, the increased signal allows us to more precisely define the required binding site composition, and second, as we discuss later, the greater sequence diversity gives us more flexibility to shuffle sites. We assembled the training set from orthologous regions in the twelve fly genomes. This poses a challenge, as the distance both makes orthology calling more difficult and opens the possibility for the function of the sequence, and hence the binding

site composition, to change. Our simulations also show that binding site turnover events can cause the enhancer to move away from its orthologous position. To address these issues, we chose to use binding site composition itself as a tool to discover orthologous enhancers' locations. Several factors are known to bind the even-skipped stripe two enhancer. These include the activators hunchback and bicoid as well as the repressors giant and Kruppel. The sequence also contains a strong binding site for the transcription factor Zelda. We know the sequence affinities of each of these transcription factors, and therefore can assess any given sequence for the number of strong and weak sites of each. While most genetic study of transcription factor binding sites in this and other enhancers has focused on the function of strong binding sites, we wished to also model the function of weak sites. We incorporated counts of sites into a chi-squared-like descriptive statistic that measured the difference between the numbers, of both strong and weak sites for these five factors, found in a target sequence and expected in a background model. By walking a window of target sequences across the eve locus, we were able to identify the location of the melanogaster stripe two enhancer. We also found a peak corresponding to the stripes four/six enhancer, which, notably, has binding sites for several of these factors as well (fig. 1).

Repeating this step in the other twelve fly genomes, we found that the same peak was visible, in approximately the same orthologous region, in flies as distantly related to melanogaster as D. willistoni. In D. anannasae, the peak was weaker, and beyond D. willistoni, the peak was not visible. The genomic sequence was missing for D. simulans, but due to its close relation to D. melanogaster, this sequence added little information in sequence diversity or binding site divergence. Thus, we assembled a training set of D. melanogaster, D. erecta, D. sechellia, D. yakuba, D. pseudoobscura, D. persimilis, and D. willistoni.

An evolutionary model allows stepwise spacing perturbation

Ideally, we would like to perturb the spatial arrangements of binding sites within the enhancer in a stepwise fashion, so that any perceived variation in expression can be tied to a small set of binding site turnover events. To this end, we simulated the evolution of the enhancer in much the same manner as described in the main text. We used a consensus approach to assemble our selective model. For each factor, we counted the number of hits that its weight matrix matched in each sequence across a wide range of cutoffs. Then, for each cutoff, we hypothesized that the minimum number of hits to this factor's binding site found across species could represent, intuitively, a minimum level of binding activity for that enhancer to function. In this manner, we populated a matrix representing a set of hypothesized selective constraints on the evolution of the enhancer sequence.

These selective constraints necessarily represent less, and sometimes substantially less, sites per sequence for each factor than is found in the typical enhancer in the training set. However, they also represent a minimum site count found in the evolutionary process: sites are gained and lost at a certain rate over time, causing the number of sites to fluctuate. These selective constraints impose a lower bound upon that fluctuation, but the typical number of sites at equilibrium is expected to be higher, and possibly sufficiently high to recreate the number of sites seen in the typical training set enhancer. To test this, we ran 100 simulations to equilibrium and counted the number of sites at each cutoff, finding that, in general, the average equlibrium number of binding sites generated by our model closely matched the average number of sites

found in the original enhancers (fig. 2).

As expected, the homology of our simulated sequences to the original D. melanogaster enhancer sequence degraded with the number of mutation-selection iterations in our model. However, this homology appeared to degrade faster in our simulations than it does between species. We evolved the D. melanogaster sequence to a distance representing, in neutral substitutions per site, the ancestor of the D. melanogaster and D. pseudoobscura, and we compared the divergence of this simulated sequence with that of the actual D. pseudoobscura sequence (fig. 3). Notably, the simulated sequence, which was created to have approximately one half of the divergence of the latter, appeared markedly more divergent.

This mismatch between simulated and real divergence can be traced to three possible causes. In the first, our neutral mutation model is miscalibrated for the divergence between the species. While it meets our needs in generating diversity for our selective model, it relies on a number of simplifying assumptions that may not hold in these species. For the second, it is possible that the enhancers are in fact tightly spatially restricted, making binding site turnover substantially more difficult and in turn slowing the pace of substitution. However, this is the very hypothesis that these simulated sequences are designed to test, so this possibility can be temporarily disregarded. Third, and finally, it is possible that we are not modeling all of the constraint placed upon the enhancer by its composition of binding sites. While the even-skipped stripe two enhancer is relatively well studied, the likelihood is low that we have a complete catalog of its bound and functionally-relevant proteins. For instance, despite decades of research into the working of the enhancer, the role of Zelda is only now being described.

Word-based scrambling preserves uncharacterized binding sites

To characterize the binding site composition of our training set without limiting ourselves to known binding sites, we turned to a word-based scrambling approach. By using the original sequences to train a Markov model, we can maintain the frequencies of words found in the set, which in turn allows us to maintain the presence of uncharacterized binding sites. We tried several versions of this model, varying the length of the words that it was trained on. For each word length W, we assembled every instance of each word of that length found in the training set. Then, for each word, we discovered the distribution of letters following that word, building a Wth-order Markov model of the sequences. In this way, by choosing a starting word at random from the set, we can easily generate new, scrambled sequences.

Choosing the optimal word length involves a tradeoff between conserving the binding site composition and shuffling the binding sites correctly. Long words will accurately represent binding sites but may only rarely be found more than once in the training set, creating long chains of unscrambled sequence. On the other hand, short words will provide a rich diversity of sequences for scrambling, but will likely only present degraded binding sites. We determined the ability of words of length five, six, seven, and eight to both scramble the sequence and preserve site composition (fig. 4). Sequences scrambled with a word length of six appeared to produce the optimal tradeoff.

Sequences chosen for testing mix Markov and evolutionary models

We chose to synthesize and test in transgenic flies two sequences from the evolutionary model and one sequence generated by a sixmer Markov model. As the sixmer model already

accomplishes a complete scrambling of the sequence, we chose the evolutionary model sequences to test more modest divergences. First, we used 181 mutation-selection rounds, roughly corresponding to the distance between D. melanogaster and its ancestor with D. erecta, to alter the original D. melanogaster enhancer. This sequence exhibited modest weakening of most of the originals' strong bicoid sites, weakening of about half of the originals' Kruppel and giant sites, and disruption of all of the original's hunchback sites. We then further altered this sequence with another 221 mutation-selection rounds. This sequence contained complete disruption of about half of all factors' strong binding sites and weakening of about half of the remainder. Analysis of the expression patterns produced by these sequences is ongoing.

Figures

Figure 1. Enhancers show changes in binding site density. Binding site density score, described in text, is plotted against the *even-skipped* locus. Location of the stripe two enhancer marked with a red asterisk.

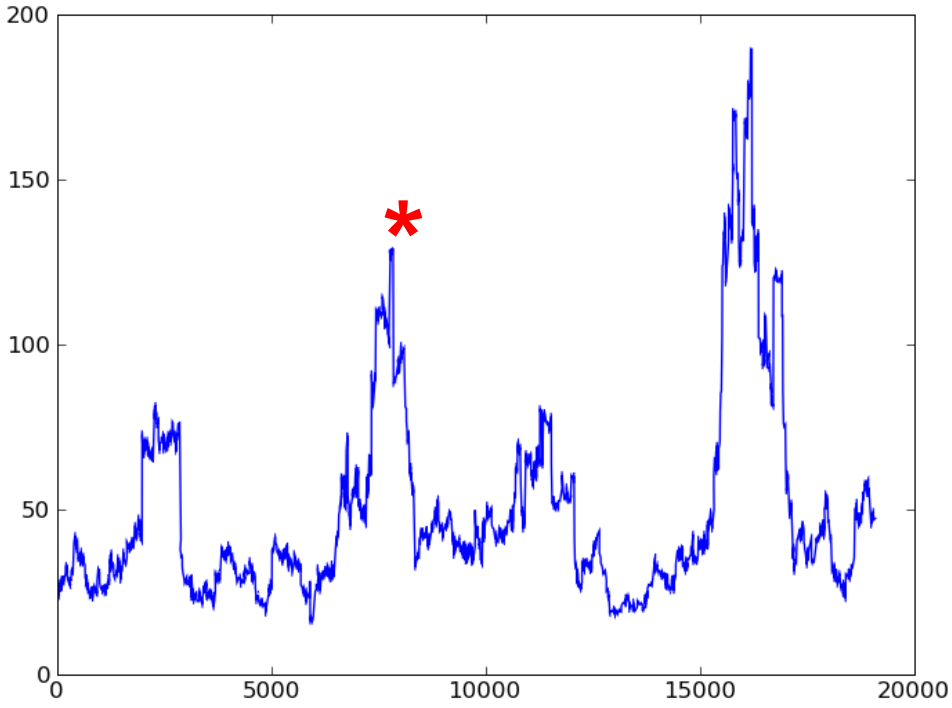


Figure 2. A consensus model reproduces typical site composition. For five factors across a range of cutoffs, the ratio of the average count of that binding site at that cutoff in simulated enhancers at equilibrium to the average count found in the training set. Color (blue: low, red: high) corresponds to the matching ratio.

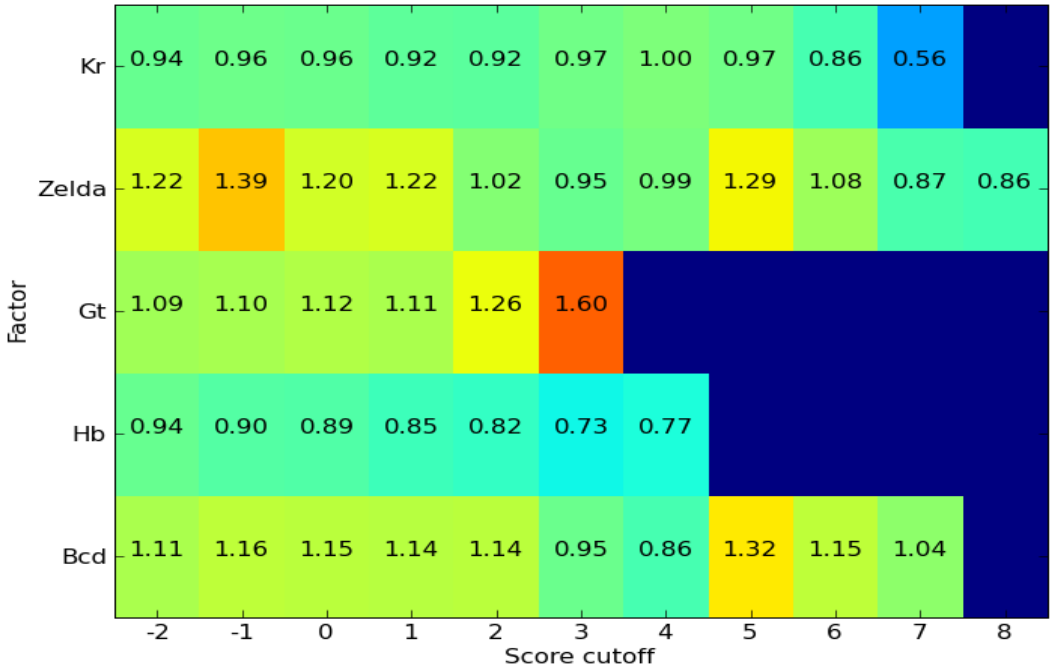
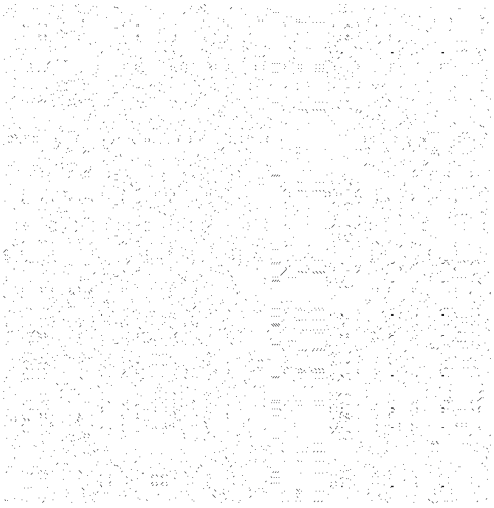


Figure 3. Rate of divergence accelerated in simulated system. Above, a dotplot comparing a sequence evolved to the distance separating D. melanogaster from its ancestor with D. pseudoobscura. Dots are plotted when matching 4mers are identical. Below, D. melanogaster vs. D. pseudoobscura.

A.



B.

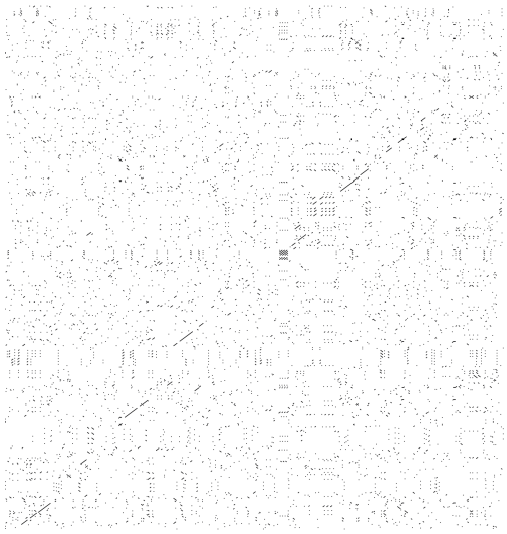
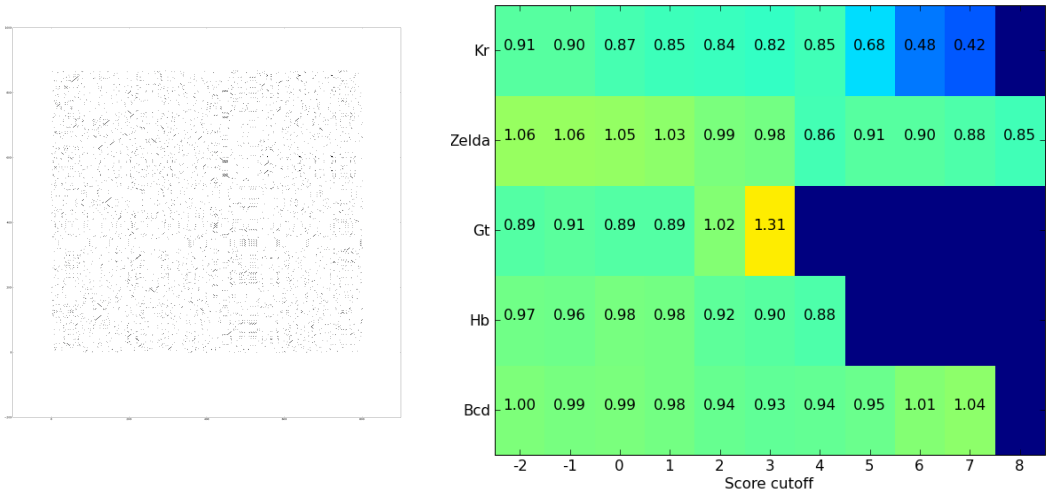


Figure 4. Tradeoffs between binding site representation and spatial scrambling. On the right, dotplots comparing a simulated sequence to the D. melanogaster sequence. On the left, heatmaps describing the conserved binding site composition of the scrambled sequences, as described in fig. 2. A and B, respectively, correspond to the 6mer and 8mer Markov models.

A.



B.