Nucleotide Interdependency in Transcription Factor Binding Sites in the *Drosophila* Genome

Jacqueline M. Dresch[1], Rowan G. Zellers[2,3], Daniel K. Bork[2,3] and Robert A. Drewell[4]

[1]Department of Mathematics and Computer Science, Clark University, Worcester, MA, USA. [2]Computer Science Department, Harvey Mudd College, Claremont, CA, USA. [3]Mathematics Department, Harvey Mudd College, Claremont, CA, USA. [4]Biology Department, Clark University, Worcester, MA, USA.

ABSTRACT: A long-standing objective in modern biology is to characterize the molecular components that drive the development of an organism. At the heart of eukaryotic development lies gene regulation. On the molecular level, much of the research in this field has focused on the binding of transcription factors (TFs) to regulatory regions in the genome known as *cis*-regulatory modules (CRMs). However, relatively little is known about the sequence-specific binding preferences of many TFs, especially with respect to the possible interdependencies between the nucleotides that make up binding sites. A particular limitation of many existing algorithms that aim to predict binding site sequences is that they do not allow for dependencies between nonadjacent nucleotides. In this study, we use a recently developed computational algorithm, MARZ, to compare binding site sequences using 32 distinct models in a systematic and unbiased approach to explore nucleotide dependencies within binding sites for 15 distinct TFs known to be critical to *Drosophila* development. Our results indicate that many of these proteins have varying levels of nucleotide interdependencies within their DNA recognition sequences, and that, in some cases, models that account for these dependencies greatly outperform traditional models that are used to predict binding sites. We also directly compare the ability of different models to identify the known KRUPPEL TF binding sites in CRMs and demonstrate that a more complex model that accounts for nucleotide interdependencies performs better when compared with simple models. This ability to identify TFs with critical nucleotide interdependencies in their binding sites will lead to a deeper understanding of how these molecular characteristics contribute to the architecture of CRMs and the precise regulation of transcription during organismal development.

KEYWORDS: transcription factor, binding site, position weight matrix, *cis*-regulatory module, *Drosophila*

CITATION: Dresch et al. Nucleotide Interdependency in Transcription Factor Binding Sites in the *Drosophila* Genome. *Gene Regulation and Systems Biology* 2016:10 21–33 doi: 10.4137/GRSB.S38462.

TYPE: Original Research

RECEIVED: February 05, 2016. **RESUBMITTED:** April 17, 2016. **ACCEPTED FOR PUBLICATION:** April 28, 2016.

ACADEMIC EDITOR: James Willey, Editor in Chief

PEER REVIEW: Three peer reviewers contributed to the peer review report. Reviewers' reports totaled 1074 words, excluding any confidential comments to the academic editor.

FUNDING: This work was funded by the National Institutes of Health (GM090167) and the National Science Foundation (IOS-0845103) grants to RAD, the National Institutes of Health (GM110571) grant to RAD and JMD, and Howard Hughes Medical Institute Undergraduate Science Education Program grants (52006301 and 52007544) to the Biology Department at the Harvey Mudd College. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jdresch@clarku.edu

Introduction

One of the largest challenges in modern biology is to understand precisely how gene expression is controlled during embryonic development. In the model organism, *Drosophila melanogaster*, as in many other species, this process is regulated at the molecular level by transcription factor (TF) proteins selectively binding to DNA sequences in *cis*-regulatory modules (CRMs) to activate or repress the transcription of target genes.[1–7] Therefore, comprehensively identifying the nature of the DNA sequences that such TFs bind to will facilitate a more complete understanding of the genetic control of development. There are several experimental techniques, including *DNase*I footprinting, protein-binding microarrays, and chromatin immunoprecipitation (ChIP),[8–11] that provide information about specific TF–DNA interactions. In general, regulatory proteins do not bind to just one DNA sequence. Though many TFs have a consensus binding site, they frequently bind to a number of other sequences as well.[12] Further complicating matters, the binding affinity between a TF and a DNA sequence motif may vary and, as a result, each distinct binding site has its own probability of being bound by a specific regulatory protein.[13]

One commonly used approach to model the preferred binding sequences for a given TF is the position weight matrix (PWM).[14] This is a matrix of size $4 \times l_w$, where l_w is the length of each of the aligned binding site sequences obtained directly from experimental data. The rows in the matrix represent each of the four possible nucleotides, A, C, G, and T, each column in the matrix represents a distinct nucleotide position in a binding site motif, and, most frequently, each element of the matrix represents the log-likelihood of observing a certain nucleotide at that position.[13–19] Unfortunately, there are several significant shortcomings of the standard PWM approach,[20–22] the most important of which may be the assumption that the frequency of a nucleotide at any given position in the matrix is independent of the nucleotide frequencies at the neighboring positions.[23] This underlying assumption of mononucleotide matrix models is particularly problematic, given that TFs can

have more than one DNA-binding domain and may be capable of contacting many nucleotides in parallel.[4,20,24–27]

There are a number of proposed methods to improve PWMs, including developing more advanced statistical algorithms for PWM construction,[21] considering higher order Markov models for the background model[28] and extending the PWM model by scoring dinucleotides and creating 16 by $l_w - 1$ matrices, analogous to mononucleotide PWMs.[14,23] For example, comparison of mononucleotide and dinucleotide weight matrices in predicting the GC-box binding motif for the human Sp1 TF revealed that the dinucleotide matrix performed with greater specificity and sensitivity.[14] A more recent study compared 26 different methods for the prediction of binding sites of 66 different TFs in mouse.[21] These methods included more sophisticated algorithms for generating simple mononucleotide PWMs and dinucleotide PWMs and performing analyses using multinucleotide (*n*-mer) matrices of arbitrary length. The results indicate that, though in many cases simple mononucleotide PWMs perform similar to more complex models, robust predictions for a subset of TFs require a more complex *n*-mer model.[21]

One important limitation of the existing *n*-mer models is that they consider groups of *n* nucleotides simultaneously, ascribing equal consideration for each nucleotide.[29–31] As a result, the predictions from such a model may be skewed by the consideration of nucleotides that are functionally irrelevant. We approach this problem by developing a new method that considers all possible gapped *n*-mer matrices,[22] including *n*-mer matrices as well as those that do not consider some of the *n* nucleotide positions. This approach allows for the possibility of ignoring several nucleotides within each frame without introducing bias into the matrices selected for analysis.[22] Our goal is to build matrices that have greater predictive efficacy than existing models without using a particularly sophisticated construction algorithm or filtered dataset. To our knowledge, the performance of such matrices has not been investigated to

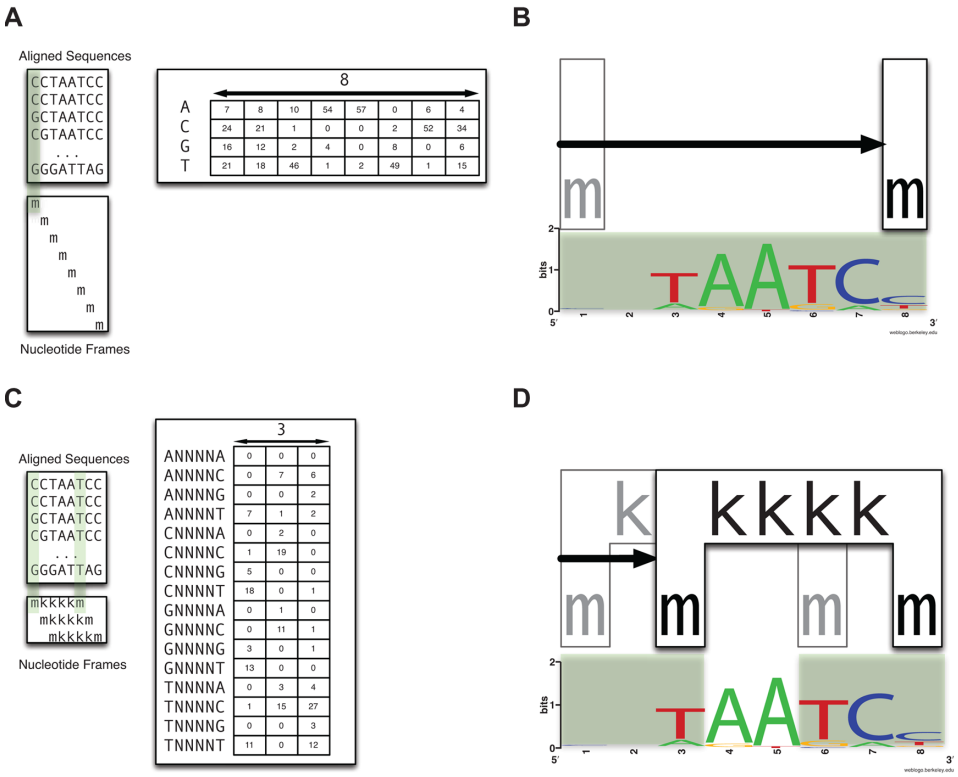

Figure 1. Comparison of a traditional mononucleotide frequency matrix and the gapped *n*-mer approach. (**A**) The traditional mononucleotide frequency matrix considers each of the eight nucleotides (*m*) in the BICOID binding site independent of the others, sliding across the sequences with a window frame size of one nucleotide. The 4×8 frequency matrix constructed contains the observed frequencies of each base over the 8 bp binding site for BICOID, obtained from *in vivo* binding data. (**B**) Visualization of the standard mononucleotide matrix approach at a BCD binding site. Highlighted nucleotides represent those that contribute to the score of each binding region. Note that each nucleotide in the 8 bp binding region contributes to the score. (**C**) The gapped *n*-mer approach uses an l_n-length window frame that considers some nucleotides (*m*) while ignoring others (*k*). The gapped *n*-mer *mkkkkm*, shown as an example, considers only the two outermost nucleotides in each frame while ignoring the inner four. Note that in this case, $l_n = 6$. This generates a 16×3 matrix, representing the frequencies of the 16 possible nucleotide pairs that can be found separated by a distance of four nucleotides within the binding motif. The three columns correspond to the three possible positions of the *mkkkkm* window sliding over the 8 bp BICOID binding region. Note that an analogous frequency matrix can be constructed from any possible gapped *n*-mer, with composition dependent on the gapped *n*-mer used. (**D**) Visualization of the gapped *n*-mer *mkkkkm* at a BCD binding site. Only the base frequencies of the highlighted nucleotides contribute to the score of each binding site. Note that the middle two nucleotides in the 8 bp binding region do not contribute to the score produced by this particular matrix.

Table 1. The 15 *Drosophila* TFs studied The TFs and their corresponding abbreviations and classifications by spatial expression profile are listed.

Transcription factor	Symbol	Classification
Bicoid	BCD	AP
Caudal	CAD	AP
Dichaete	D	AP
Dorsal	DL	DV
Fushi-Tarazu	FTZ	PR
Giant	GT	AP
Hairy	H	PR
Hunchback	HB	AP
Huckebein	HKB	AP
Knirps	KNI	AP
Kruppel	KR	AP
Paired	PRD	PR
Runt	RUN	PR
Schnurri	SHN	DV
Sloppy Paired 1	SLP	PR

Abbreviations: AP, anterior–posterior; DV, dorsal–ventral; PR, pair-rule.

date. Herein, we present a systematic evaluation of 32 distinct gapped *n*-mer matrices for modeling the DNA sequence specificity of 15 different TFs that regulate gene expression in early *Drosophila* development. Our results show that many of these regulatory proteins have their own preferred set of gapped *n*-mer matrices based on sequence specificity and that in many cases these gapped *n*-mer matrices perform better than simple *n*-mers or mononucleotide PWMs. In addition, two distinct groups of TFs are identified, which demonstrate radically different scoring profiles.

Results

Experimental approach. To investigate the ability of distinct matrix models to predict the DNA sequence specificity of *Drosophila* TFs, we employ our previously developed MARZ algorithm and the associated RZ scoring method.[22] The MARZ algorithm utilizes a systematically constructed set of 32 matrices to perform an unbiased analysis of TF binding sequences. The matrices include the simple mononucleotide model, *m* (Fig. 1A and B), and all possible gapped *n*-mer matrices with a reading frame of length less than or equal to six nucleotides. The gapped *n*-mer matrices only consider a subset of nucleotides (represented by an *m*) and ignore the other nucleotides (represented by a *k*) in the frame. For example, the *mkkkkm* matrix considers only the two outside nucleotides in a six-nucleotide frame (Fig. 1C and D). In total, 15 different *Drosophila* TFs were analyzed, each of which is known to play an important role in regulating early embryonic development (Table 1). The experimentally determined binding sites for these 15 TFs range in size from 7 to 15 bp and are variable in sequence, resulting in PWMs with different overall information content (Fig. 2).

When determining whether a potential binding site scores high enough to be defined by MARZ as a predicted binding site, the algorithm uses a *threshold position*, *x* (ranging from 0 to 1). The scoring threshold is then determined by identifying the highest threshold at which 1 – *x* percentage of the experimentally determined binding sites are identified as

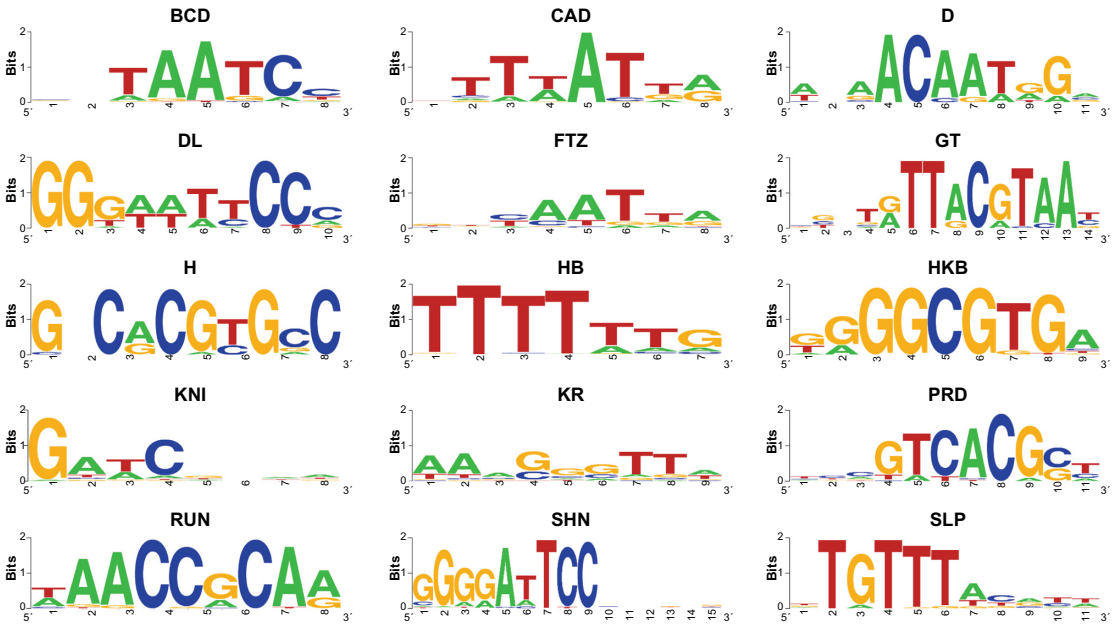

Figure 2. Sequence logos for the 15 TFs in this study. Sequence logos are a visualization tool to represent the information content of each nucleotide at each position in a (mononucleotide) PWM for a particular TF. The sequence logos were constructed using the software developed by Crooks et al.[58] and experimentally verified binding sites[45,56,57] as described in the "Methods" section.

the predicted binding sites. Thus, a high threshold position only allows for the prediction of strong binding sites, while a low threshold position allows for the prediction of a wide range of weak and strong binding sites.

MARZ analysis reveals two distinct groups of TFs. The RZ score measures the ability of a matrix model to correctly predict the genuine binding sites identified in ChIP peaks taken from the *Drosophila* genome.[22] For the 15 TFs analyzed, two distinct scoring profiles are observed (Fig. 3). Ten of the TFs (D, DL, GT, H, HB, HKB, PRD, RUN, SHN, and SLP – Group A) show very similar profiles with a number of shared features, as follows. (i) The matrix scores monotonically decrease as the threshold position increases. (ii) There is a very narrow range of scores observed (0.50–0.55) for almost all TFs across all thresholds. (iii) Matrix performance tends to be more variable at lower threshold positions and becomes more uniform as the threshold position increases. In contrast, the five other TFs (BCD, CAD, FTZ, KNI, and KR – Group B) show scoring profiles that do not contain the shared Group A-identified features and are also quite distinct from one another. Comparison of the known expression patterns in the embryo, the regulatory network to which they belong, or the characterized DNA-binding domain(s) in each of the TFs in either Group A or B fails to identify any obvious biological motivation for this observed grouping based solely on these criteria (Table 2).

Table 2. Grouping of the 15 *Drosophila* TFs by overall matrix performance.

TF	Classification	DNA binding domain
Group A		
D	AP	HMG box
DL	DV	RHD
GT	AP	bZip2
H	PR	HLH
HB	AP	ZF-C2H2
HKB	AP	ZF-C2H2
PRD	PR	Paired Homeodomain
RUN	PR	Runt
SHN	DV	ZF-C2H2
SLP	PR	Fork head
Group B		
BCD	AP	Homeodomain
CAD	AP	Homeodomain
FTZ	PR	Homeodomain
KNI	AP	ZF-C4
KR	AP	ZF-C2H2

Note: The TFs are listed and classified by their spatial expression profile and the identity of their primary DNA-binding domain. The TFs fall into two distinct groups when the RZ scores of all 32 matrices are analyzed, as shown in Figure 3. Group A consists of 10 TFs that show a relatively narrow range of values following a simple pattern of monotonically decreasing RZ scores with increasing threshold value. In contrast, TFs in Group B (BCD, CAD, FTZ, KNI, and KR) show more complex patterns across the range of thresholds.
Abbreviations: AP, anterior–posterior; DV, dorsal–ventral; PR, pair-rule.

Group A TFs – shared common features. It is clear that the scoring profiles of Group A TFs share a number of key features. The simple mononucleotide matrix, *m*, performs relatively well for the majority of TFs (rank #1–5), with the exception of H and HB (Table 3). In addition, the dinucleotide matrix, *mm*, performs worse than the *m* matrix, except in the case of H, HB, and SLP. Matrices containing strings of ungapped nucleotides (eg, *mmm*, *mmmm*) or containing few gaps relative to total matrix length (eg, *mmkmmm*, *mmmmkm*) also perform relatively poorly (Table 3). In contrast, some of the *n*-mer matrices containing many gaps relative to matrix length consistently perform well for this group of TFs, notably, *mkkm*, *mkkkm*, and *mkkkkm*. However, it should be noted that for all Group A TFs, the narrow range of overall scores (0.50–0.55) across all thresholds signifies that all of the 32 different matrix models score similarly (Fig. 3), suggesting that their predictive power is also quite similar. Taken together, these features indicate that in most cases for these 10 TFs matrices with fewer nucleotide interdependencies appear to perform just as well as, or better than, more complicated matrices (Table 3). If we only examine the top five scoring matrices for each TF, we see a clear repetition of the pattern observed across all 32 matrices, namely, a monotonic decrease in the predictive score as the threshold increases, within a relatively narrow overall range of scores (Fig. 4).

These results suggest that the TFs in Group A have similar preferences for nucleotide interdependencies in their respective binding site sequences. In fact, the relative strength of the performance of the *m* matrix, combined with the weaker performance of the *mm* and longer strings of ungapped *m* matrices, suggests little interdependency between the individual adjacent nucleotides in the binding sites. The inability of the vast majority of the 32 matrix models to outperform the simple mononucleotide *m* model supports this conclusion and indicates that there may be limited value in using any of the more complex matrix models to predict *in vivo* binding sites for this set of TFs.

Group B TFs – distinct scoring profiles. Unlike the relatively uniform scoring profiles shared by all of the Group A TFs, the profiles for the five TFs in Group B are much more distinct. We no longer observe the threshold-dependent monotonic decrease in RZ scores that characterizes the Group A TFs, but rather each of the Group B factors exhibit somewhat distinctly shaped profiles (Figs. 3 and 4). Despite these individual differences (which are discussed in more detail below), a few general trends are observed for the Group B TFs. (i) There is greater variability in scores from different matrix models at lower threshold positions, which is reduced at higher thresholds. (ii) Greater overall variability in scores is observed for many of the Group B matrices, when compared with Group A. (iii) The overall range of scores (0.30–0.66) is much larger

Table 3. Heat map ranking each matrix/TF combination by RZ score.

		Group A TFs										Group B TFs					Mean			
Type ID	Gapped *n*-mer	D	DL	GT	H	HB	HKB	PRD	RUN	SHN	SLP	BCD	CAD	FTZ	KNI	KR	Group A	Group B	ALL TFs	Ratio A/B
0	m	2	5	1	22	28	2	4	1	5	2	6	28	21	20	4	7.2	15.8	10.1	0.46
1	mm	13	18	4	10	19	25	13	5	15	1	11	19	15	25	16	12.3	17.2	13.9	0.72
2	mkm	10	1	2	11	15	18	7	7	13	8	19	16	23	26	21	9.2	21	13.1	0.44
3	mmm	24	9	12	14	17	30	21	10	19	15	25	13	17	11	13	17.1	15.8	16.7	1.08
4	mkkm	4	3	13	2	12	3	2	2	4	3	4	30	1	27	25	4.8	17.4	9.0	0.28
5	mkmm	15	11	22	4	8	10	14	13	10	13	7	12	2	15	20	12	11.2	11.7	1.07
6	mmkm	21	12	5	1	4	26	12	8	12	9	30	23	22	16	23	11	22.8	14.9	0.48
7	mmmm	29	10	23	7	14	19	26	18	24	25	24	7	18	4	12	19.5	13	17.3	1.50
8	mkkkm	1	2	6	13	9	5	1	4	11	5	12	29	5	21	15	5.7	16.4	9.3	0.35
9	mkkmm	6	14	14	15	31	11	9	11	7	6	1	32	25	18	22	12.4	19.6	14.8	0.63
10	mkmkm	3	16	7	3	11	12	5	9	8	16	28	14	19	10	18	9	17.8	11.9	0.51
11	mkmmm	18	23	24	19	10	6	19	16	26	22	9	20	30	19	19	18.3	19.4	18.7	0.94
12	mmkkm	9	27	15	6	16	23	8	15	14	7	14	26	9	5	26	14	16	14.7	0.88
13	mmkmm	20	30	16	16	32	15	22	21	22	18	13	9	28	2	14	21.2	13.2	18.5	1.61
14	mmmkm	25	13	25	12	6	20	20	19	25	24	27	21	26	8	24	18.9	21.2	19.7	0.89
15	mmmmm	31	21	26	20	7	21	27	27	30	27	16	11	31	30	27	23.7	23	23.5	1.03
16	mkkkkm	5	4	8	23	1	1	3	6	2	4	22	1	32	32	2	5.7	17.8	9.7	0.32
17	mkkkmm	7	7	9	28	18	13	11	12	1	10	20	25	16	24	3	11.6	17.6	13.6	0.66
18	mkkmkm	11	8	17	8	24	16	6	14	6	12	8	27	20	17	11	12.2	16.6	13.7	0.73
19	mkkmmm	16	15	10	25	20	7	17	26	16	19	2	31	6	12	7	17.1	11.6	15.3	1.47
20	mkmkkm	12	6	3	17	3	4	10	3	3	14	18	10	7	13	9	7.5	11.4	8.8	0.66
21	mkmkmm	19	22	11	27	23	8	15	22	18	26	3	5	24	3	5	19.1	8	15.4	2.39
22	mkmmkm	23	19	27	9	26	14	16	20	17	23	23	17	12	1	17	19.4	14	17.6	1.39
23	mkmmmm	28	29	28	26	27	9	24	24	28	30	5	4	27	14	31	25.3	16.2	22.3	1.56
24	mmkkkm	8	20	18	21	2	31	18	17	9	11	32	3	29	7	10	15.5	16.2	15.7	0.96
25	mmkkmm	22	24	19	29	13	27	25	28	21	21	10	15	8	22	1	22.9	11.2	19.0	2.04
26	mmkmkm	14	26	20	5	21	32	23	25	23	20	26	8	4	23	6	20.9	13.4	18.4	1.56
27	mmkmmm	26	25	21	31	22	22	28	29	31	31	15	2	11	28	29	26.6	17	23.4	1.56
28	mmmkkm	17	17	29	24	5	28	29	23	20	17	29	24	3	6	8	20.9	14	18.6	1.49
29	mmmkmm	30	31	30	32	25	24	30	31	29	28	17	18	10	31	30	29	21.2	26.4	1.37
30	mmmmkm	27	28	31	18	29	29	31	30	27	29	31	22	14	9	28	27.9	20.8	25.5	1.34
31	mmmmmm	32	32	32	30	30	17	32	32	32	32	21	6	13	29	32	30.1	20.2	26.8	1.49

Notes: This chart shows the ranking of all gapped *n*-mer matrices for each TF. This is done by ordering the RZ scores at their peak thresholds (see "Methods" section). Herein, a score of 1 indicates the best performing gapped *n*-mer for the TF; a score of 32 indicates the worst performing gapped *n*-mer for that TF. These are color coded such that green (lower) entries indicate relatively strong performance, whereas red (higher) entries indicate relatively poor matrix performance. The average rank of each gapped *n*-mer matrix across all TFs in each of Groups A and B, as well as over all 15 TFs, are also presented. The ratio of the average rank of each matrix across group A to its average rank across Group B is also shown; in this section, blue (lower) entries denote a stronger performance across Group A, while white (higher) entries indicate stronger performance across Group B.

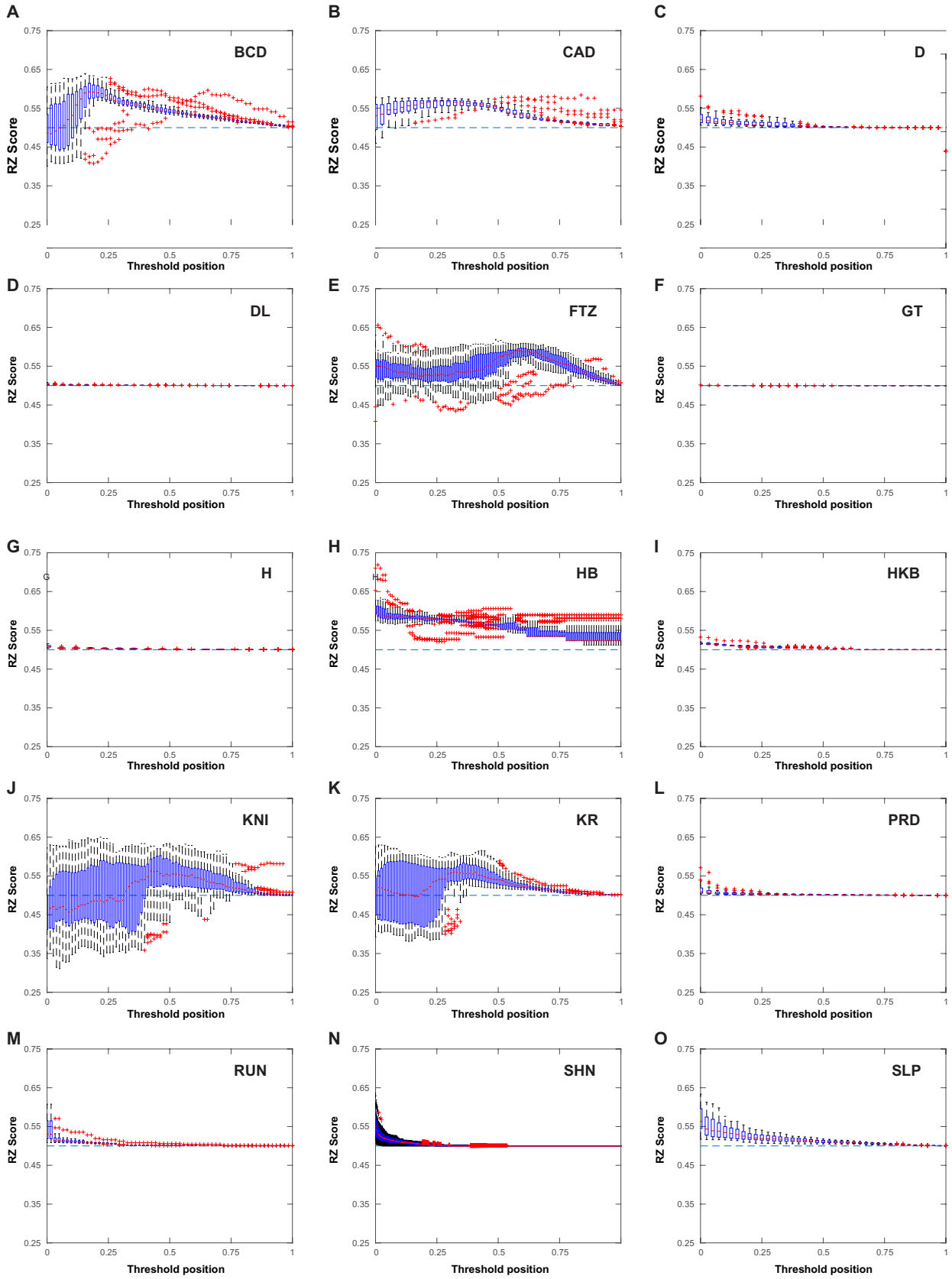

Figure 3. Comparing the performance of the 32 gapped *n*-mer matrices for all 15 TFs. In each graph, the *x*-axis corresponds to the threshold position used for each run of the MARZ algorithm. The *y*-axis corresponds to the RZ score obtained from each run (see "Methods" section for details). At a given threshold, the central mark (a red line) represents the median RZ score of the 32 gapped *n*-mer matrices, the blue boxes enclose the 25th to 75th percentiles of the data set, the whiskers extend to all other points not considered outliers, and the outliers are plotted separately (red crosses). In random simulations, the RZ scores obtained are approximately 0.5.[22] Thus, we have included a dashed line at 0.5, representing the score obtained in the case of nondiscrimination.

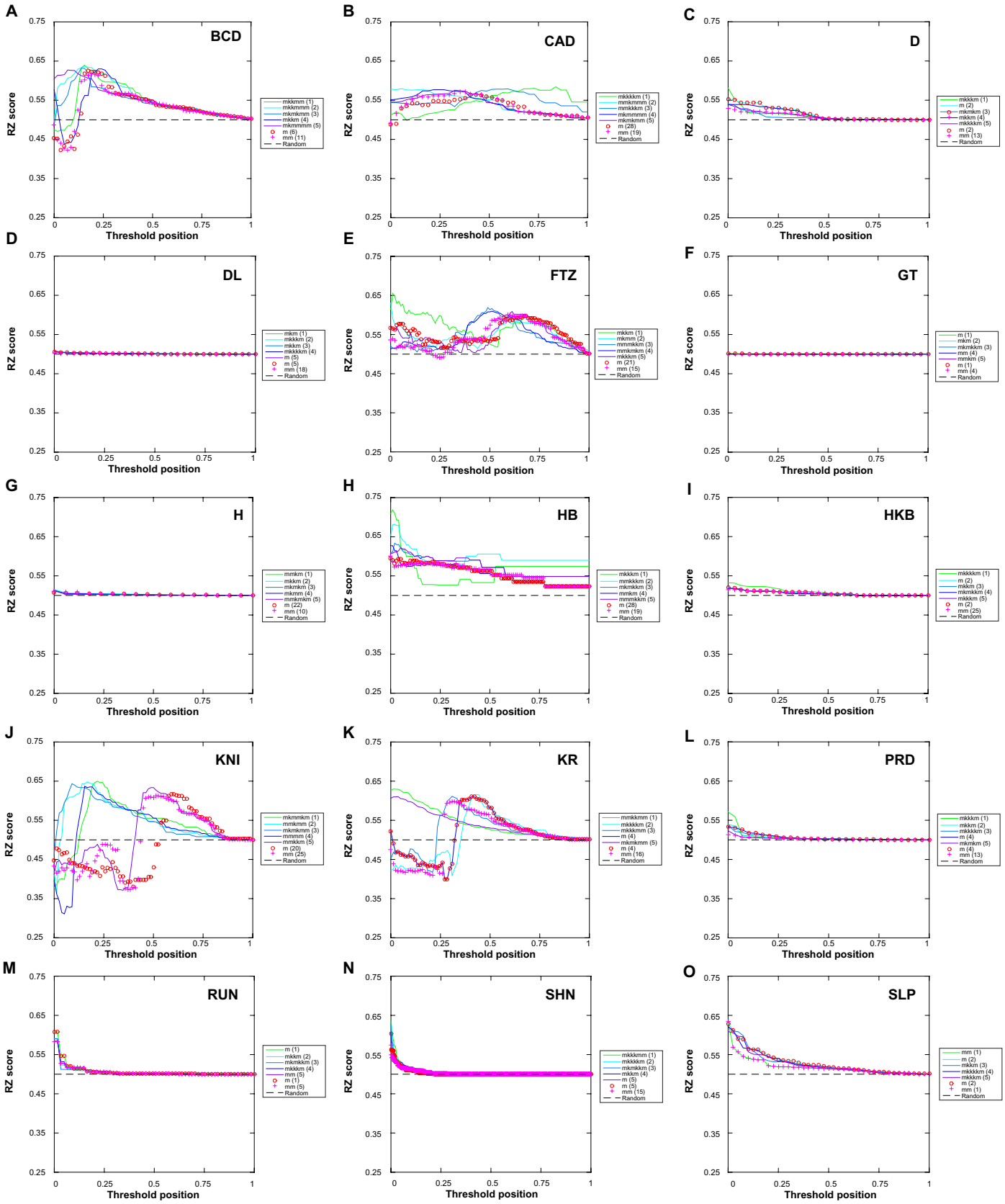

Figure 4. Performance of the top five matrices over the full range of thresholds for all 15 TFs. In each graph, the *x*-axis corresponds to the threshold position used for each run of the MARZ algorithm. The *y*-axis corresponds to the RZ score obtained for each run (see "Methods" section for details). The RZ scores of the five highest-scoring matrices are shown, with the rank of each matrix indicated in parentheses. The standard mononucleotide (*m*) and dinucleotide (*mm*) matrices are also included for comparison. For many of the TFs, all matrices consistently outperform the no-discrimination case (dashed line), especially at low threshold values. The top matrices for a number of TFs (see BCD, CAD, FTZ, KNI, and KR) show complex performance variation with threshold position.

than the range found in the Group A TFs. (iv) Performance at any given threshold varies considerably with the matrix type. (v) Performance of each matrix changes in response to the threshold used (although often in a unique fashion). Given the radically different scoring profiles for each of the five Group B TFs (Fig. 4), we will discuss the details of each TF profile individually.

BICOID. BCD is a homeodomain TF that, in common with other homeodomain-containing TFs, is thought to bind strongly to the canonical TAAT recognition sequence[32–34] observed in the center of its 8 bp PWM (Fig. 2). Therefore, it is a strong candidate for nucleotide interdependencies within its binding site sequences. This is supported by the fact that some of the best performing matrices for BCD are models with extended gaps, including *mkkmm*, *mkkmmm*, and *mkkm* (Table 3) that exhibit nonadjacent nucleotide dependencies. However, even these top-ranked matrices do not perform uniformly well over all threshold values (Fig. 4), emphasizing the need for careful consideration of the stringency at which predictive analyses are run. In addition, it is clear that longer nongapped matrices, including *mmm* and *mmmm*, which are ranked 25th and 24th, respectively, out of the 32 different models, do not perform well for BCD (Table 3). This may indicate that extended stretches of adjacent nucleotide dependency are not a feature of BCD binding sites.

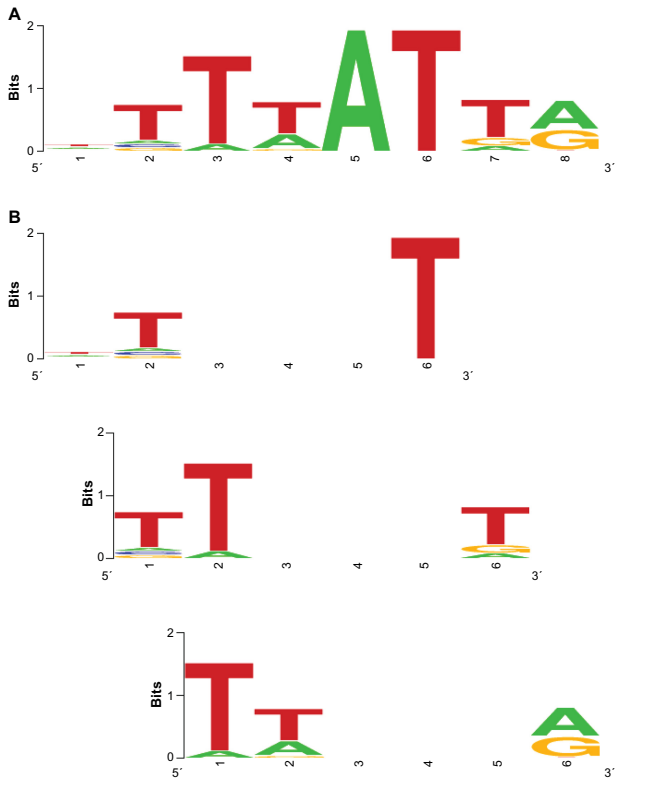

Figure 5. Visualizing the nucleotides included in a gapped *n*-mer model. (**A**) A standard sequence logo for CAUDAL, corresponding to the nucleotide frequencies in a traditional mononucleotide model, constructed using the software developed by Crooks et al.[58] and experimentally verified binding sites[45,56,57] as described in the "Methods" section. (**B**) A series of sequence logos to represent the nucleotides considered in each sliding window used with the best performing model, *mkkkkm*, to predict 8 bp binding site for CAUDAL.

CAUDAL. The CAD TF also contains a DNA-binding homeodomain[25,35] with a relatively robust T(A/T)AT motif represented in the center of its 8 bp PWM (Fig. 2). In the case of CAD, the top five ranked gapped *n*-mers all contain six positions (Fig. 4). Noticeably, the mononucleotide *m* model performs very poorly (ranked 28th out of 32), while *mmmm*, *mmmmm*, and *mmmmmm* are all ranked in the top 11 models (Table 3). Together, these results may be indicative of a wide amount of nucleotide dependence in CAD binding sites. If we only consider *n*-mer models that contain gapped (*k*) positions, an interesting additional observation about CAD is revealed. The *mmkkkm* model is ranked 3rd, while its reflection *mkkkmm* is 25th (Table 3). This large difference in predictive ability implies that the orientation of gaps considered in the binding sites is significant. The actual nucleotides being considered when using the *mmkkkm* model are visualized in Figure 5. A further feature of the CAD matrices is that they perform relatively consistently across the entire range of threshold values (Fig. 3) when compared with the other TFs in Group B.

FUSHI-TARAZU. FTZ is the third homeodomain TF[36,37] in Group B and also has a characteristic (T/C)AAT motif represented in the center of its 8 bp PWM (Fig. 2). The profile for FTZ shows a relatively large range of RZ scores (0.42–0.66), which is generally larger in the lower half of threshold values and decreases at higher thresholds (Fig. 3). The *m* matrix performs poorly (ranked 21st), while gapped *n*-mer matrices with nonadjacent nucleotide dependencies perform the best: *mkkm*, *mkmm*, and *mmmkkm* are ranked 1st, 2nd, and 3rd, respectively (Table 3). These results suggest that, like the other homeodomain TFs, nonadjacent nucleotide dependencies are important in FTZ binding sites. However, the exact arrangement of gapped positions considered in the binding site also appears to be critical, as the *mkkkkm* model performs the worst of all 32 matrices (Table 3).

KNIRPS. KNI contains a characterized C4-type zinc finger (ZF) DNA-binding domain,[38,39] and its corresponding PWM is unusual in that it contains four adjacent nucleotide positions with high information content accompanied by four nucleotides with very little information content (Fig. 2). This may, in part, be the result of KNI binding to multiple distinct motifs.[20] Therefore, it is perhaps not surprising that all of the top five performing matrix models for KNI contain at least four nucleotide positions [one has four positions (*mmmm*), two have five positions (*mmkmm* and *mmkkm*), and two have six positions (*mkmmkm* and *mkmkmm*); Table 3]. Of these five matrices, the top three ranked models all contain an *mkm* motif within them, suggesting that nonadjacent nucleotide dependencies are present in the KNI binding sites (Fig. 2). This is supported by the fact that the dinucleotide model *mm*, which only considers adjacent dependencies, performs very poorly (ranked 25th). Of note is the very broad range of scores

(0.32–0.65) for the top five ranked models across different thresholds (Fig. 4), once more suggesting that the stringency at which the analysis is carried out is critically important to the predictive ability of different matrix models.

KRUPPEL. KR has a C2H2-type ZF DNA-binding domain.[40,41] Its PWM represents a 9 bp binding site with relatively low information content at all positions when compared with many of the other TFs in our study (Fig. 2). In the case of KR, the simple mononucleotide model *m* is ranked 4th overall, suggesting that nucleotide interdependencies may not be prevalent in its binding sites. However, the scoring profile for KR shows some subtleties that may underlie a more complex DNA-protein binding regime. All of the top five ranked models (with the exception of the mononucleotide matrix) contain six nucleotide positions (Table 3) and the top three all have nonadjacent dependencies separated by a string of *k*'s (*mmkkmm*, *mkkkkm,* and *mkkkmm*). As was the case for KNI, a large range of scores are seen in the overall profile (0.31–0.64), with the general trend of a wider range of scores at low thresholds and a corresponding narrowing of score range at higher thresholds (Figs. 3 and 4).

Identification of known KRUPPEL binding sites in CRMs. To directly compare the predictive ability of a gapped model to the simple mononucleotide and dinucleotide models, we investigated the identification of experimentally verified KR binding sites by the top-ranked *mmkkmm* matrix in three different well-characterized *Drosophila* CRMs, namely, *even-skipped* stripe 2 enhancer (S2E[42]), *Abdominal-B* IAB5 enhancer (IAB5[43]), and *Abdominal-B* IAB7b enhancer (IAB7b[44]). In each case, the ability of the different models to rigorously identify the known functional KR binding sites (true positives) and avoid false-positive predictions using the lowest possible threshold value (see "Methods" section) is assessed (Table 4). The S2E contains three known KR binding sites, which are successfully identified by the *m* and *mm* models. However, at this permissive threshold value, these models also identify 630 and 274 false-positive binding sites on the two DNA strands in the 480 bp enhancer, respectively (Table 4, S2E). In contrast, the *mmkkmm* matrix only identifies 17 false-positive binding sites, but does fail to identify one of the known KR binding sites. The difference in the performance of the models is also illustrated by the specificity (or true-negative rate), calculated as TN/(TN + FP) (where TN is true negatives and FP is false positives), of these three matrices (*m:* 0.336; *mm:* 0.711; and *mmkkmm:* 0.982). The IAB5 and IAB7b enhancers both contain two known KR binding sites, which are successfully identified by all the three models. Once more, the false positive rate is drastically reduced and the specificity is greatly increased (*m:* 0.316 and 0.352; *mm:* 0.692 and 0.690; and *mmkkmm:* 0.996 and 0.997, respectively) for both enhancers when the *mmkkmm* model is applied when compared with the simple models (Table 4, IAB5 and IAB7b), indicating that the gapped model is outperforming the *m* and *mm* models at this threshold.

Performing the same analysis at the threshold values at which the *m*, *mm*, and *mmkkmm* models received their highest RZ score (see "Methods" section) reveals similar results (Supplementary Table 1), albeit the false-positive rate is reduced for all three models. This result once again emphasizes that the stringency at which the analysis is carried out is critically important to the predictive ability of different matrix models.

Discussion

The results of our study demonstrate that, in order to increase the accuracy of predicting *in vivo* binding sites for transcription factors (TFs), it is critical to carefully consider which gapped *n*-mer models to employ and the threshold level at which the analysis is performed. In many cases, complex gapped *n*-mer matrices outperform traditional simple mononucleotide or *n*-mer models. In addition, two distinct groups of TFs are identified with radically different scoring profiles, suggesting that optimal model selection may depend on TF-specific protein-DNA interaction regimes.

Advantage of using the MARZ algorithm. In comparison to a number of previous bioinformatics tools used for the identification of transcription factor binding sites (12, 14–19, 21, 23), the MARZ algorithm offers one clear advantage; a systematic and unbiased consideration of all gapped and non-gapped nucleotide dependence. MARZ enables the user to compare the performance of complex gapped *n*-mer matrices to traditional simple mononucleotide or *n*-mer models and reports a clear scoring profile (Table 3). In addition, the algorithm systematically tests all models and therefore also allows for the detection of cases where simple, more traditional matrices outperform more complex gapped matrices.

TF profiles fall into two distinct groups. The scoring profile for all five of the Group B TFs (BCD, CAD, FTZ, KNI, and KR) differs substantially from the shared profile features of TFs in Group A (D, DL, GT, H, HB, HKB, PRD, RUN, SHN, and SLP; summarized in Fig. 3 and Table 3). By considering the ratio between each matrix model's average ranked performance on the Group A TFs and the Group B TFs, we directly compare the overall performance of each individual model (Table 5). Accordingly, a ratio <1 indicates that the model performs better on Group A TFs and a ratio >1 indicates that the model performs better on Group B TFs (Table 5). Using this metric enables us to identify some key patterns. First, simpler matrices, represented by independent nucleotides (including the mononucleotide matrix) and more gaps, more accurately predict the binding sites for Group A TFs. Correspondingly, more complex matrices, represented by higher numbers of nucleotide interdependencies, are more accurate predictors of binding sites for Group B TFs (Table 5). This indicates that the binding sites for the Group A TFs predominantly contain arrangements of nucleotides that are independent of each other and that utilizing complex matrix models to search for binding sites may be unnecessary in these cases. In contrast, many of the complex gapped *n*-mer models perform

Table 4. KRUPPEL binding site predictions in CRMs.

Threshold	low		
Matrix	**m**	**mm**	**mmkkmm**
S2E			
TP	3	3	2
FP	630	274	17
FN	0	0	1
TN	319	675	932
IAB5			
TP	2	2	2
FP	569	256	3
FN	0	0	0
TN	263	576	829
IAB7b			
TP	2	2	2
FP	188	90	1
FN	0	0	0
TN	102	200	289

Notes: The ability of the mononucleotide (*m*), dinucleotide (*mm*), and gapped (*mmkkmm*) matrix models to identify the known KR binding sites at the lowest threshold value. Predicted binding sites in each of the three CRMs are classified as true positives (TP), false positives (FP), false negatives (FN), or true negatives (TN). In each case, the performance of the gapped matrix is compared with the mononucleotide matrix (red, performs worse; blue, performs equally; green, performs better). For all three CRMs, the complex gapped matrix predicts many fewer false-positive sites.
Abbreviations: *even-skipped* stripe 2 enhancer (S2E), *Abdominal-B* IAB5 enhancer (IAB5), and *Abdominal-B* IAB7b enhancer (IAB7b).

relatively well for Group B TFs, suggesting that the binding sites for these proteins do in fact harbor a number of critical nucleotide interdependencies. However, it should be noted that the predictive ability of individual models tends not to be systematic across the five Group B TFs (Table 5), emphasizing, emphasizing the need for careful analysis of any particular TF under investigation.

Although we have included TFs with a wide variety of different DNA-binding domains (Table 2), it is of note that all three homeodomain-containing proteins (BCD, CAD, and FTZ) included in our analysis fall into Group B. In the case of each of these TFs, the homeodomain is thought to be largely responsible for mediating protein–DNA interactions,[32,45] predominantly via contact with an evolutionarily conserved core consensus TAAT motif.[32,33,35–37] This motif is represented in the center nucleotide positions (3–6) of the 8 bp PWMs for all three TFs in our study (Fig. 2). Given the demonstrated importance of this TAAT binding sequence, it is therefore reasonable to expect some degree of dependence among nucleotides within 3 bp of one another in the binding sites for BCD, CAD, and FTZ. Such a hypothesis is supported by the fact that for all three TFs, complex matrix types, that consider many nucleotide positions, demonstrate relatively strong predictive performance (Table 3). Intriguingly, the gapped matrices that perform well for these three homeodomain TFs may also offer some insight into the physical mechanisms underlying the nucleotide interdependencies in their DNA-binding sites and the associated binding affinity with the protein TF. All three TFs are thought to not only mediate binding interactions in the major groove of DNA using the homeodomain recognition helix but also additional DNA contact in the minor groove via the relatively unstructured N-terminal domain.[32,36,37,46] Therefore, it is possible that when considering binding site specificity, the identity of the nucleotides in the transition positions between the major and minor groove may not be as important as the nucleotides in the grooves themselves. There is support for this idea from the fact that the best performing matrices for each of the three TFs contain extended nucleotide gaps, namely, *mkkmm* (BCD), *mkkkkm* (CAD), and *mkkm* (FTZ) (Table 3).

A complex gapped matrix model outperforms traditional models for KRUPPEL. Direct comparison of the predictive ability of the top-ranked gapped *n*-mer model (*mmkkmm*) with the traditional mononucleotide (*m*) and dinucleotide (*mm*) models reveals that the gapped model is much more selective at identifying the known KR binding sites in three well-characterized *Drosophila* CRMs (Table 4). In the case of the *even-skipped* stripe 2 enhancer (S2E), the gapped model predicts over an order of magnitude fewer false-positive binding sites when compared with the two simple models (Table 4) at the lowest threshold value. Intriguingly, this selectivity also results in the mis-identification of one of the three known KR sites as a false negative, suggesting that there may be a potential trade-off in the predictive ability. In the case of the IAB5 and IAB7b enhancers, all three models correctly identify the known KR binding sites, but again the gapped model greatly outperforms the two traditional models in terms of predicting fewer false-positive sites (Table 4). Comparison of the model predictions at the threshold values at which the *m*, *mm*, and *mmkkmm* matrices receive their highest RZ score reveals a similar pattern of results (Supplementary Table 1). However, in these cases, the false-positive rate is reduced for all three models, indicating that the threshold stringency at which the analysis is performed can influence the predictive ability of different matrix models.

Conclusions

Given the considerable difficulty in developing reliable methods to identify TF binding sites in complex metazoan genomes, the MARZ algorithm will be a useful addition to the currently available repertoire of bioinformatics tools. Of note is the similarity of our results with a previous study investigating the performance of different predictive models for mammalian TFs.[21] In the earlier study, 26 models were applied to predict binding sites for 66 different mouse TFs. The results indicated that models based on simple mononucleotide PWMs perform similarly to more complex models for most of the mouse TFs examined, but do not perform as well in some cases (<10% of the TFs examined).[21] In our current study, we find an analogous

Table 5. Heat map ranking each matrix/TF Group combination by average RZ score.

		Mean		
Type ID	Gapped *n*-mer	Group A	Group B	Ratio A/B
4	mkkm	4.8	17.4	0.28
16	mkkkkm	5.7	17.8	0.32
8	mkkkm	5.7	16.4	0.35
2	mkm	9.2	21	0.44
0	m	7.2	15.8	0.46
6	mmkm	11	22.8	0.48
10	mkmkm	9	17.8	0.51
9	mkkmm	12.4	19.6	0.63
20	mkmkkm	7.5	11.4	0.66
17	mkkkmm	11.6	17.6	0.66
1	mm	12.3	17.2	0.72
18	mkkmkm	12.2	16.6	0.73
12	mmkkm	14	16	0.88
14	mmmkm	18.9	21.2	0.89
11	mkmmm	18.3	19.4	0.94
24	mmkkkm	15.5	16.2	0.96
15	mmmmm	23.7	23	1.03
5	mkmm	12	11.2	1.07
3	mmm	17.1	15.8	1.08
30	mmmmkm	27.9	20.8	1.34
29	mmmkmm	29	21.2	1.37
22	mkmmkm	19.4	14	1.39
19	mkkmmm	17.1	11.6	1.47
31	mmmmmm	30.1	20.2	1.49
28	mmmkkm	20.9	14	1.49
7	mmmm	19.5	13	1.50
26	mmkmkm	20.9	13.4	1.56
23	mkmmmm	25.3	16.2	1.56
27	mmkmmm	26.6	17	1.56
13	mmkmm	21.2	13.2	1.61
25	mmkkmm	22.9	11.2	2.04
21	mkmkmm	19.1	8	2.39

Notes: This chart shows the average ranks of each gapped n-mer matrix across all TFs in each of Groups A and B. These are color-coded such that green (lower) entries indicate relatively strong performance, whereas red (higher) entries indicate relatively poor matrix performance. The ratio of the average rank of each matrix across group A to its average rank across group B is also shown; in this section, blue (lower) entries denote a stronger performance across Group A, while white (higher) entries indicate stronger performance across Group B. Gapped n-mers have been ordered according to their ratio of average rank across Group A to its average rank across Group B.

situation – for the 10 Group A *Drosophila* TFs, there is very little difference between the performance of the 32 different models in the MARZ algorithm. However, for the five Group B *Drosophila* TFs, the best performing models are those that include complex gapped matrices with nucleotide interdependencies. Investigation of the underlying biological implications of the performance of these different models will be critical in future studies. In particular, expanding the studies to examine the performance of the MARZ algorithm on additional datasets for Group B TF binding, including *in vivo* ChIP-seq[47] and *in vitro* PBMs,[9] will be informative in further dissecting the key nucleotide interdependencies in the binding sites. Given that the sequences in ChIP peaks often contain binding sites for multiple TFs,[48,49] it may also be possible that some of the interdependencies detected by the MARZ algorithm represent overlapping sequences from two different binding sites. In future studies, it will therefore be important to explore other salient features of TF binding sites in CRMs,[50] including their spatial arrangement,[51–53] relative binding affinities,[8,30,54] and the biophysical constraints of protein–DNA interactions,[55] in combination with the application of gapped *n*-mer matrix models, in order to further refine overall predictive efficacy.

Methods

Data used. In this study, we investigate 15 TFs prominent in embryonic *Drosophila* development. These were chosen based on the degree of their characterization, the quality and quantity of the corresponding data, and the range of their spatial expression profiles (Table 1).

For each TF, ChIP-chip data were obtained from MacArthur et al.[49], corresponding to regions of DNA in which the given TF binds. To reduce any potential noise in the data, only the center 100 bp of each ChIP peak are considered. Any ChIP peak of fewer than 100 bp of length is discarded, thus all trimmed ChIP peaks used are exactly 100 bp in size. These trimmed ChIP peaks, combined with aligned sequences from *in vivo* binding data,[45,56,57] are then used as the input to the MARZ algorithm.[22]

MARZ algorithm. The MARZ algorithm combinatorially analyzes all possible gapped *n*-mer matrices (where $n \leq 6$) for each studied TF. A gapped *n*-mer is defined as a string of *k*'s and *m*'s such that any nucleotides located at an *m* contribute to the score of the potential binding sequence and are therefore assumed to depend on one another (see Fig. 1 and Ref. 22 for details). Due to the varying length and composition among the gapped *n*-mers, each encodes different assumptions about nucleotide dependence/independence in putative binding sites. The MARZ algorithm as described in the study by Zellers et al.[22] can thus be used to determine which assumptions are most suitable for analyzing a given TF at a given threshold value.

Threshold value. When analyzing a given TF, the MARZ algorithm identifies as binding sites those that score above a given threshold. To determine a threshold, *T*, a threshold position, *x* (ranging from 0 to 1), is used. This corresponds to the highest threshold at which the best-scoring $1 - x$ percentage of aligned sequences are all identified as binding sites. Thus, a high threshold position only allows for the prediction of strong binding sites, while a low threshold position allows for the prediction of a wide range of weak and strong binding sites.

Weight matrices and RZ scores. MARZ constructs a weight matrix from aligned binding sequence data by

constructing first a frequency matrix from the aligned sequences, and then comparing this with a background matrix constructed from the entire *Drosophila* genome. Similar to how traditional PWMs are constructed, we then construct a weight matrix with columns that represent the contributions of each possible nucleotide composition of the gapped *n*-mer at each sliding frame along a potential binding site. Using this matrix, if a string of nucleotides is scored greater than or equal to the threshold *T*, it is predicted to be a binding site.

The RZ score is a measure of the effectiveness of a gapped *n*-mer weight matrix at a given threshold. It is a value in the range [0, 1] that corresponds to the ability of a weight matrix to differentiate real from scrambled ChIP peaks at a given threshold. Note that if the number of predicted binding sites on a true ChIP peak is greater than the average number of predicted binding sites on the corresponding scrambled ChIP peaks, 1 point is added to the score. If the average number of predicted binding sites on the scrambled ChIP peaks is greater than the number of predicted binding sites on the corresponding true ChIP peak, 0 is added to the score. If scrambled and true ChIP peaks have exactly the same number of predicted sites, 0.5 is added to the score. This score is then divided by the number of ChIP peaks to give a measure of effectiveness in the range [0, 1]. An overall value of 0.5 corresponds to a matrix that offers no predictive power.[22]

For each TF, MARZ computes the RZ scores over all gapped *n*-mers and threshold positions. We also provide a ranking of all gapped *n*-mers for a given TF. Since gapped *n*-mers exhibit varying performance over different thresholds, we order the gapped *n*-mers used to predict binding sites for a given TF according to their peak RZ score over all threshold positions. The raw RZ scores used to assemble Table 3 can be found in Supplementary Table 2.

KRUPPEL binding site predictions. The minimal CRM DNA sequences and the location of experimentally verified KRUPPEL binding sites were obtained from previous studies for *even-skipped* stripe 2 enhancer (480 bp S2E[42]), *Abdominal-B* IAB5 enhancer (425 bp cIAB5[43]), and *Abdominal-B* IAB7b enhancer (154 bp 2F2K region of IAB7b[44]).

Statement on Data and Regent Availability

The MARZ algorithm code is available upon request.

Acknowledgment

We thank Lily Li for contributing to the computational analysis in this study.

Author Contributions

Conceived the study, performed the computational analysis, and drafted the manuscript: RGZ. Participated in the design of the study and statistical analysis and drafted the manuscript: DKB. Conceived the study, participated in its design and coordination, and drafted the manuscript: RAD, JMD. All the authors read and approved the final manuscript.

Supplementary Materials

Supplementary Table 1. KRUPPEL binding site predictions in CRMs at different threshold values

Supplementary Table 2. Raw RZ score data used to assemble Table 3.

REFERENCES

1. Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Development*. 1997;124(10):1851–64.
2. Badis G, Berger MF, Philippakis AA, et al. Diversity and complexity in DNA recognition by transcription factors. *Science*. 2009;324:1720–3.
3. Davidson EH. A view from the genome: spatial control of transcription in sea urchin development. *Curr Opin Genet Dev*. 1999;9:530–41.
4. Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*. 2004;116:247–57.
5. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*. 1989;245:371–8.
6. Ptashne M. Gene regulation by proteins acting nearby and at a distance. *Nature*. 1986;6081:697–701.
7. Ptashne M, Gann A. Transcriptional activation by recruitment. *Nature*. 1997;6625:569–77.
8. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc*. 2009;4(3):393–411.
9. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*. 2015;43(Database issue):D117–22.
10. Whitington T, Frith MC, Johnson J, Bailey TL. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res*. 2011;39:e98.
11. Zhu LJ, Christensen RG, Kazemian M, et al. FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res*. 2011;39:D111–7.
12. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*. 2000;16(1):16–23.
13. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. 1999;15:563–77.
14. Gershenzon NI, Stormo GD, Ioshikhes IP. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic Acids Res*. 2005;33:2290–301.
15. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37:W202–8.
16. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins: statistical-mechanical theory and application to operators and promoters. *J Mol Biol*. 1987;193:723–43.
17. Djordjevic M, Sengupta AM, Shraiman BI. A biophysical approach to transcription factor binding site discovery. *Genome Res*. 2003;13:2381–90.
18. Hertz GZ, Hartzell GW III, Stormo GD. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci*. 1990;6:81–92.
19. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res*. 2005;33:5781–98.
20. Stringham JL, Brown AS, Drewell RA, Dresch JM. Flanking sequence context-dependent transcription factor binding in early *Drosophila* development. *BMC Bioinformatics*. 2013;14:298.
21. Weirauch MT, Cote A, Norel R, et al; DREAM5 Consortium. Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol*. 2013;31(2):126–34.
22. Zellers RG, Drewell RA, Dresch JM. MARZ: an algorithm to combinatorially analyze gapped n-mer models of transcription factor binding. *BMC Bioinformatics*. 2015;16:30.
23. Siddharthan R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS One*. 2010;5(3):e9722.
24. Jolma A, Yan J, Whitington T, et al. DNA-binding specificities of human transcription factors. *Cell*. 2013;152:327–39.
25. Mlodzik M, Fjose A, Gehring WJ. Isolation of caudal, a *Drosophila* homeo box-containing gene with maternal expression, whose transcripts form a concentration gradient at the pre-blastoderm stage. *EMBO J*. 1985;4:2961–9.
26. Rothe M, Nauber U, Jäckle H. Three hormone receptor-like *Drosophila* genes encode an identical DNA-binding finger. *EMBO J*. 1989;8:3087–94.
27. Sommer RJ, Retzlaff M, Goerlich K, Sander K, Tautz D. Evolutionary conservation pattern of zinc-finger domains of *Drosophila* segmentation genes. *Proc Natl Acad Sci U S A*. 1992;89:10782–6.

28. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc*. 2008;3(10):1578–88.
29. Kulakovskiy I, Levitsky V, Oshchepkov D, Bryzgalov L, Vorontsov I, Makeev V. From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J Bioinform Comput Biol*. 2013;11(1):1340004.
30. Mathelier A, Wasserman WW. The next generation of transcription factor binding site prediction. *PLoS Comput Biol*. 2013;9(9):e1003214.
31. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordan R. Stability selection for regression-based models of transcription factor-DNA binding specificity. *Bioinformatics*. 2013;29(13):i117–25.
32. Gehring WJ, Qian YQ, Billeter M, et al. Homeodomain-DNA recognition. *Cell*. 1994;78(2):211–23.
33. Hanes SD, Brent R. DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell*. 1989;57(7):1275–83.
34. Struhl G, Struhl K, Macdonald PM. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell*. 1989;57(7):1259–73.
35. Berger MF, Badis G, Gehrke AR, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*. 2008;133(7):1266–76.
36. Desplan C, Theis J, O'Farrell PH. The sequence specificity of homeodomain-DNA interaction. *Cell*. 1988;54(7):1081–90.
37. Schier AF, Gehring WJ. Direct homeodomain-DNA interaction in the autoregulation of the fushi tarazu gene. *Nature*. 1992;356(6372):804–7.
38. Arnosti DN, Gray S, Barolo S, Zhou J, Levine M. The gap protein knirps mediates both quenching and direct repression in the *Drosophila* embryo. *EMBO J*. 1996;15(14):3659–66.
39. Nauber U, Pankratz MJ, Kienlin A, Seifert E, Klemm U, Jäckle H. Abdominal segmentation of the *Drosophila* embryo requires a hormone receptor-like protein encoded by the gap gene knirps. *Nature*. 1988;336(6198):489–92.
40. Schuh R, Aicher W, Gaul U, et al. A conserved family of nuclear proteins containing structural elements of the finger protein encoded by Kruppel, a *Drosophila* segmentation gene. *Cell*. 1986;47(6):1025–32.
41. Stanojevic D, Hoey T, Levine M. Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in *Drosophila*. *Nature*. 1989;341(6240):331–5.
42. Small S, Blair A, Levine M. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. *EMBO J*. 1992;11(11):4047–57.
43. Starr MO, Ho MC, Gunther EJ, et al. Molecular dissection of cis-regulatory modules at the *Drosophila* bithorax complex reveals critical transcription factor signature motifs. *Dev Biol*. 2011;359:290–302.
44. Drewell RA, Nevarez MJ, Kurata JS, Winkler LN, Li L, Dresch JM. Deciphering the combinatorial architecture of a *Drosophila* homeotic gene enhancer. *Mech Dev*. 2014;131:68–77.
45. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*. 2008;133(7):1277–89.
46. Baird-Titus JM, Clark-Baldwin K, Dave V, Caperelli CA, Ma J, Rance M. The solution structure of the native K50 Bicoid homeodomain bound to the consensus TAATCC DNA-binding site. *J Mol Biol*. 2006;356(5):1137–51.
47. Bradley RK, Li XY, Trapnell C, et al. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol*. 2010;8(3):e1000343.
48. Li XY, MacArthur S, Bourgon R, et al. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol*. 2008;6:e27.
49. MacArthur S, Li XY, Li J, et al. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol*. 2009;10:R80.
50. Gallo SM, Gerrard DT, Miner D, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Res*. 2011;21:456–64.
51. Cha M, Zhou Q. Detecting clustering and ordering binding patterns among transcription factors via point process models. *Bioinformatics*. 2014;30(16):2263–71.
52. Ng FS, Schütte J, Ruau D, et al. Constrained transcription factor spacing is prevalent and important for transcriptional control of mouse blood cells. *Nucleic Acids Res*. 2014;42(22):13513–24.
53. Ochoa-Espinosa A, Yucel G, Kaplan L, et al. The role of binding site cluster strength in Bicoid-dependent patterning in *Drosophila*. *Proc Natl Acad Sci U S A*. 2005;102(14):4960–5.
54. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the *in vivo* effects of binding site variants. *Genome Res*. 2013;23(11):1908–15.
55. Haldane A, Manhart M, Morozov AV. Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol*. 2014;10(7):e1003683.
56. Ho MC, Johnsen H, Goetz SE, et al. Functional evolution of cis-regulatory modules at a homeotic gene in *Drosophila*. *PLoS Genet*. 2009;5:e1000709.
57. Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Res*. 2008;36(8):2547–60.
58. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–90.